International Review of the Red Cross (2024), **106** (926), 732–759.
Selected Articles
doi:10.1017/S1816383124000067

IRRC

Is wearing these sunglasses an attack? Obligations under IHL related to anti-AI countermeasures

Jonathan Kwik*
Postdoctoral Researcher, ELSA Lab Defence, T. M. C. Asser Institute, The Hague, Netherlands
Email: j.kwik@asser.nl, j.h.c.kwik@gmail.com

Abstract

As usage of military artificial intelligence (AI) expands, so will anti-AI countermeasures, known as adversarials. International humanitarian law offers many protections through its obligations in attack, but the nature of adversarials generates ambiguity regarding which party (system user or opponent) should incur attacker responsibilities. This article offers a cognitive framework for legally analyzing adversarials. It explores the technical, tactical and legal dimensions of adversarials, and proposes a model based on foreseeable harm to determine when legal responsibility should transfer to the countermeasure's author. The article provides illumination to the future combatant who ponders, before putting on their adversarial sunglasses: "Am I conducting an attack?"

Keywords: countermeasure, artificial intelligence, autonomous weapon, precautions in attack, precautions in defence, adversarial, poisoning, backdoor.

*ORCID No. 0000-0003-0367-5655. The author thanks Dr William Boothby for his helpful comments and insights during the drafting of this paper.
The advice, opinions and statements contained in this article are those of the author/s and do not necessarily reflect the views of the ICRC. The ICRC does not necessarily represent or endorse the accuracy or reliability of any advice, opinion, statement or other information provided in this article.

Introduction

In 2017, Massachusetts Institute of Technology (MIT) researchers 3D-printed an animal inside a hard round shell, with four protruding limbs, a head and a tail – to any incidental observer, clearly a turtle. However, when photos of the object were fed into Google's image classifier, the output was disconcerting: "rifle". Even when the object was rotated, flipped and turned, the algorithm remained convinced: "rifle", "rifle", "rifle", predicted with around 90% certainty.[1] The team had successfully created a "robust adversarial example": an undetectable pattern that could fool an otherwise well-performing artificial intelligence (AI) algorithm into consistently misclassifying the object as something else.

The experiment brought much attention to the fallibility of modern algorithms and their vulnerability to targeted manipulation by enemies, and anti-AI countermeasures have been mentioned as a major concern by many military organizations.[2] Such countermeasures also potentially place the civilian population at an aggravated risk: if an opposing force were to cause an autonomous military system to classify handbags as rifles near a busy market, in order to disperse fire away from its own soldiers and towards market-goers, the civilian toll could be immense.

This article explores **adversarials**[3] – AI-specific countermeasures – and how applicable rules related to distinction and precautions should be interpreted when adversarials are employed. In particular, the matter will be considered from an *interactive* perspective, inviting into the discussion both the system's owner, hereinafter referred to as the **deployer**, and the countermeasure's author, hereinafter referred to as the **adversary**.[4] The aim is to understand different types of adversarials and to construct a framework for assessing them in a legal sense, allowing us to determine how international humanitarian law (IHL) would apply. Removing the legal fog related to their use would benefit the protection of civilians and improve accountability for both deployers and adversaries.

Let us start by putting adversarials in context. In terms of warfighting, the basic intent behind adversarials is nothing new. Belligerents have employed

1 See Anish Athalye, Logan Engstrom, Andrew Ilyas and Kevin Kwok, *Synthesizing Robust Adversarial Examples*, 6th International Conference on Learning Representations, Vancouver, 2018, para. 1, available at: https://openreview.net/forum?id=BJDH5M-AW (all internet references were accessed in February 2024).

2 See e.g. US Department of Defense (DoD), "Autonomy in Weapon Systems", DoD Directive 3000.09, 25 January 2023, p. 4 (US); NATO, *Summary of the NATO Artificial Intelligence Strategy*, 22 October 2021, para. 14, available at: www.nato.int/cps/en/natohq/official_texts_187617.htm (NATO); Josh Baughman, "China's ChatGPT War", China Aerospace Studies Institute, 2023, p. 7, available at: www.airuniversity.af.edu/CASI/Display/Article/3498584/chinas-chatgpt-war/ (China).

3 The term "adversarial" is used to distinguish countermeasures specifically designed to exploit modern AI from countermeasures in general.

4 This article deliberately uses legally neutral terms instead of "attacker" and "defender" since, as argued below, these statuses may shift depending on the circumstances. The current nomenclature offers more consistency, since the deployer is always the party activating the system, and the adversary is always the author of the countermeasure.

smokescreens to hamper the accuracy of laser-guided munitions; used thermal camouflage to avoid detection by infrared cameras; used jammers to degrade radar and GPS effectiveness; and hacked enemy systems to render them inoperable.[5] These are all countermeasures against enemy technology, and it is unsurprising that the equivalent would also emerge for AI systems. Although certain methods to counter enemy operations such as perfidy and human shielding are prohibited in IHL, countermeasures generally constitute lawful ruses of war.[6] Deception, ruses and the degradation of enemy capabilities are as old as warfare itself,[7] and countermeasures in no way reduce the attacking party's duties under IHL to observe obligations related to discrimination and precautions.[8]

With the spread of hacking and cyber attacks in modern warfare, the possibility of "wresting control" of enemy systems has become more prominent. Instead of having to physically seize an opponent's capability, hackers can remotely overtake control of a system, with potentially disastrous consequences. Commentators have frequently evoked this concern with respect to remote-controlled or autonomous technologies.[9] As Scharre remarks, "a hacker could … usurp control of an autonomous weapon and redirect it",[10] potentially to harm civilians or even commit fratricide.[11] As discussed later in this article, in this scenario, the legal characterization is not particularly complicated: the seized weapon is being used to effectuate an attack by its new controller (the hacker), with all IHL rules related to attacks being applicable to that controller.[12]

With AI-specific countermeasures like the adversarial turtle, however, the legal characterization becomes more complex. Who is in control of the weapon if it temporarily misclassifies an object due to an adversarial pattern, such as a seemingly innocuous pair of sunglasses?[13] Are the sunglasses themselves an

5 Paul Scharre, *Army of None: Autonomous Weapons and the Future of War*, W. W. Norton & Co, New York, 2018, p. 40; Tetyana Krupiy, "A Case against Relying Solely on Intelligence, Surveillance and Reconnaissance Technology to Identify Proposed Targets", *Journal of Conflict and Security Law*, Vol. 20, No. 3, 2015, p. 4.

6 Protocol Additional (I) to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts, 1125 UNTS 3, 8 June 1977 (entered into force 7 December 1978) (AP I), Art. 37(2). See also William H. Boothby, "Highly Automated and Autonomous Technologies", in William H. Boothby (ed.), *New Technologies and the Law in War and Peace*, Cambridge University Press, Cambridge, 2018, p. 154.

7 Milan N. Vego, "Operational Deception in the Information Age", *Joint Forces Quarterly*, Spring 2002, p. 60.

8 Samuel Estreicher, "Privileging Asymmetric Warfare? Part I: Defender Duties under International Humanitarian Law", *Chicago Journal of International Law*, Vol. 11, No. 2, 2011, p. 435; William H. Boothby, "Control in Weapons Law", in Rogier Bartels *et al.* (eds), *Military Operations and the Notion of Control Under International Law*, T. M. C. Asser Press, The Hague, 2021, p. 384.

9 Christof Heyns, *Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions*, UN Doc. A/HRC/23/47, 9 April 2013, para. 98; Michael N. Schmitt and Jeffrey Thurnher, ""Out of the Loop": Autonomous Weapon Systems and the Law of Armed Conflict", *Harvard Law School National Security Journal*, Vol. 4, 2013, p. 242.

10 P. Scharre, above note 5, p. 222.

11 Defense Science Board, *Report of the Defense Science Board Summer Study on Autonomy*, Secretary of Defense for Acquisition, Technology and Logistics, Washington, DC, 2016, p. 14.

12 Yoram Dinstein and Arne Willy Dahl, *Oslo Manual on Select Topics of the Law of Armed Conflict*, Springer, Cham, 2020 (Oslo Manual), p. 41.

13 T. Krupiy, above note 5, p. 33.

attack, a means to effectuate an attack through an opponent's system, or "just" a countermeasure? Can the system's owner reasonably be expected to account for the consequences, when this pattern is (often designed to be) imperceptible to humans? If a belligerent compromises an AI system's training data, are they "in control" throughout the system's lifespan? Many AI-specific countermeasures blur the notions of control and foreseeability – notions which are crucial for determining which party is responsible for implementing many duties related to distinction and precautions. In the worst situation, legal uncertainty may lead parties to accuse each other of having been the "attacker", allowing both to deny responsibility for the resultant harm. Ultimately, the civilian population would pay the price. As Schmitt has argued with regard to cyber attacks,

> uncertainty with respect to the loss of functionality threshold leaves the legal characterization of certain cyber operations directed at or affecting the civilian population ambiguous. A party to the conflict could exploit such uncertainty …. From a humanitarian perspective, this situation is untenable.[14]

AI in the military domain has been the subject of legal debate for more than a decade. Most discussion has focused on the systems' *users* – more specifically, under what circumstances (if at all) AI can lawfully be used for military decision-making or to effectuate kinetic attacks.[15] While cyber attacks, hacking and other forms of interference are often raised,[16] and AI-specific vulnerabilities such as data reliance, out-of-distribution data and brittleness are sometimes referenced,[17] these matters are usually presented as issues that the systems' *owners* should account for. Relatively minimal attention has been placed on the *interfering party* and how IHL applies to its activities. The present article aims to fill this conceptual gap.

To properly analyze how to view adversarials in a legal sense, it is crucial to obtain a general understanding of how adversarials function, what vulnerabilities of modern AI they exploit, and the extent to which they can influence a target system's

14 Michael N. Schmitt, "Wired Warfare 3.0: Protecting the Civilian Population during Cyber Operations", *International Review of the Red Cross*, Vol. 101, No. 910, 2019, p. 340.

15 For a summary of the polemic regarding AI in military systems, see Elvira Rosert and Frank Sauer, "How (Not) to Stop the Killer Robots: A Comparative Analysis of Humanitarian Disarmament Campaign Strategies", *Contemporary Security Policy*, Vol. 42, No. 1, 2021, pp. 18–21.

16 See e.g. GGE on LAWS, *Report of the 2019 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*, UN Doc CCW/GGE1/2019/3, 25 September 2019, Annex IV(f); Office of the Assistant Secretary of Defense for Research and Engineering, *Technical Assessment: Autonomy*, DoD, Washington, DC, 2015 (DoD Autonomy Assessment), p. 3. For AI systems more generally, see David Leslie, *Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector*, Alan Turing Institute, London, 2019, p. 31.

17 See e.g. Defense Innovation Board, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense Defense Innovation Board*, DoD, Washington, D.C., 2019, p. 16, available at: https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF; International Committee of the Red Cross (ICRC), *Artificial Intelligence and Machine Learning in Armed Conflict: A Human-Centred Approach*, Geneva, 2019, p. 11; Ministère des Armées, *L'intelligence artificielle au service de la défense*, Ministère des Armées, Paris, 2019, p. 7.

behaviour. Thus, the article first engages in an exploration of adversarials in a more technical sense, focusing in particular on two common types: adversarial inputs and poisoning attacks. The article then dives into the tactical realm, and theorizes why an adversary would prefer some adversarials above others. Next, the article discusses the legal criteria for an action to be considered an attack, the obligations attached to both attack and "sub-attack" situations, and recent legal theories surrounding countermeasures against more autonomous technologies.

From this baseline, the article then argues for an approach that uses the foreseeability of harmful consequences to determine (1) whether an adversarial is an attack, and (2) whether the main responsibility for enacting precautionary measures should remain with the deployer or should transfer to the adversary. The article concludes with overall thoughts and recommendations for future discussions.

Adversarials from a technical lens

Many techniques exist that are tailor-made for compromising AI systems.[18] For the purposes of this article, we shall limit our discussion to **adversarial inputs** and **poisoning attacks**. The reason for focusing on these two forms of adversarials is twofold. First, they are unique to the AI domain. While viruses, cyber interference or integrity attacks[19] can be used as countermeasures against AI systems, they can equally be used against other digital software or drones; in contrast, adversarial inputs and poisoning attacks are "uniquely AI". Second, they are some of the most common adversarials,[20] and in any case are very likely to be used against military systems.[21] As shown below, adversarial inputs in particular require no access to the system, and can be used by adversaries just by manipulating elements of the operational environment. Understanding their legal status in IHL is therefore of crucial importance.

Vulnerabilities of modern AI

To address why adversarials are sufficiently different from other countermeasures to warrant narrow legal inspection, we should observe first that *any* countermeasure exploits a vulnerability in the target system. Smokescreens exploit the fact that laser-guided bombs require line-of-sight, and jamming exploits drones' reliance on communication signals. What distinguishes the adversarials discussed in this

18 See Ram Shankar Siva Kumar *et al.*, "Adversarial Machine Learning – Industry Perspectives", 2020, p. 2, available at: http://arxiv.org/abs/2002.05646; Paul Scharre, *Autonomous Weapons and Operational Risk*, Center for a New American Security, Washington, DC, 2016, pp. 34–36.

19 Integrity attacks may nonetheless be *preparatory* for a poisoning attack or model replacement.

20 Comiter similarly divides adversarials broadly into adversarial inputs and poisoning. Marcus Comiter, *Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do about It*, Belfer Center for Science and International Affairs, Cambridge, 2019, pp. 7–10.

21 P. Scharre, above note 5, p. 182.

article from other countermeasures is that they exploit a unique characteristic of AI: data reliance. Modern AI has reached the ubiquity and success that it currently enjoys due to the development of machine learning (ML), a technique allowing developers to produce high-performance systems using vast amounts of data.[22] ML has allowed breakthroughs in domains where manual programming was too onerous, too complex or simply impossible.[23] Instead of defining symbolically what a tank looks like, a system can be "taught" to recognize a tank by feeding it hundreds of pictures of tanks.[24] Given the complexity of many military tasks, ML (either in pure or hybrid form) is indispensable for many modern military AI systems.[25]

ML's source of strength, however, is also its weakness. Because an ML model's "knowledge" is directly tied to the training dataset, faults, biases and inaccuracies in the data are also transferred to the final product. "[T]he intelligence is in the data, not the algorithm", and both quality and quantity matter.[26] Ideally, training datasets should be complete, relevant, accurate, high-quality and faithful to the actual operational environment.[27] Unfortunately, the perfect dataset is very hard to compile. Data necessary to create a robust algorithm is often hard to obtain[28] or monopolized.[29] For military applications, the problem is compounded. Many pre-labelled open-source pictures are available of "dogs", but few of "infantry fighting vehicles". This limitation has been raised in both official reports and by commentators.[30] As the International Committee of the Red Cross (ICRC) has noted, modern AI's data reliance is likely "a significant issue for applications in armed conflict, where high-quality, representative data for specific tasks is scarce".[31]

Shortcuts are available to circumvent these challenges, but are not necessarily secure. Some developers have uncritically relied on open-source

22 Cristoph Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, Leanpub, 2019, p. 10; UK House of Lords, Select Committee on Artificial Intelligence, *Report of Session 2017–19: AI in the UK: Ready, Willing, and Able?*, HL Paper No. 100, 16 April 2018, p. 14.

23 Bathaee analogizes the situation with teaching a child how to ride a bike: "Although one can explain the process descriptively or even provide detailed steps, that information is unlikely to help someone who has never ridden one before to balance on two wheels. One learns to ride a bike by attempting to do so over and over again and develops an intuitive understanding." Yavar Bathaee, "The Artificial Intelligence Black Box and the Failure of Intent and Causation", *Harvard Journal of Law and Technology*, Vol. 31, No. 2, 2018, p. 902.

24 See e.g. Fan-jie Meng *et al.*, "Visual-Simulation Region Proposal and Generative Adversarial Network Based Ground Military Target Recognition", *Defence Technology*, Vol. 18, No. 11, 2022.

25 US Air Force Office of the Chief Scientist, *Autonomous Horizons: System Autonomy in the Air Force – A Path to the Future*, Vol. 1: *Human-Autonomy Teaming*, AF/ST TR 15-01, 1 June 2015, p. 22.

26 Brian Haugh, David Sparrow and David Tate, *The Status of Test, Evaluation, Verification, and Validation (TEV&V) of Autonomous Systems*, Institute for Defense Analyses, Alexandria, 2018, pp. 2–3.

27 Arthur Holland-Michel, *Known Unknowns: Data Issues and Military Autonomous Systems*, United Nations Institute for Disarmament Research (UNIDIR), Geneva, 2021, p. 1, available at: https://unidir.org/publication/known-unknowns.

28 Vincent Boulanin, *Mapping the Development of Autonomy in Weapon Systems: A Primer on Autonomy*, Stockholm International Peace Research Institute, Stockholm, 2016, p. 22.

29 UK House of Lords, above note 22, p. 45.

30 See e.g. Ministère des Armées, above note 17, p. 5; A. Holland-Michel, above note 27, p. 27.

31 ICRC, above note 17, p. 11.

datasets without verifying their source or safety,[32] or have used dragnets (where "everything that can be collected is collected").[33] Because raw data needs to be pre-labelled before it can be used, developers sometimes rely on non-experts[34] or even outsource the process.[35] In addition, training more complex ML models (e.g. convolutional neural nets) is computationally intensive. Instead of building models from scratch in-house, developers often resort to federative learning (training "smaller" models on user devices before combining the models together),[36] transfer learning (taking an existing model and fine-tuning it for a new task)[37] or outsourcing.[38] The main problem with all of these shortcuts is *exposure*: they invite outside influence into the dataset, labels or sub-models, and this influence can be ill-intentioned. While this vulnerability is often invoked in reports,[39] one would be mistaken in thinking that military developers are immune to it: even the Pentagon has been using gig-economy labellers to develop drone capabilities.[40]

A second major vulnerability derived from data reliance is the fact that ML models – despite very high average performance – can make strange, unintuitive mistakes when exposed to particular inputs. Perturbations or distortions in the input data, even minimal ones, can drastically reduce a model's accuracy: even random movements in the background can sometimes lead to unexpected errors.[41] In itself, it is not strange that input perturbations reduce accuracy: a forward observer, too, has a greater chance of mistaking a tractor for a tank during a blizzard than on a sunny day. What is unique about AI is that the patterns liable for causing such errors can be entirely unintuitive for humans. The model might not be fazed by the heaviest snowstorm, but may glitch fatally due to a random bumper sticker. The exact reason for such behaviour is outside the scope of this article,[42] but the underlying problem lies in the fact that ML relies

32 A survey found that many organizations neglect extra security testing and verification, assuming instead that available datasets are safe and tested against adversarials. Evidently, this is not the case. R. S. S. Kumar *et al.*, above note 18, p. 3.

33 M. Comiter, above note 20, p. 29.

34 This is potentially problematic for legal labels, e.g. letting laypeople label whether a person is a combatant or not.

35 *Ibid.*, p. 38.

36 See Shiqi Shen, Shruti Tople and Prateek Saxena, "Auror: Defending against Poisoning Attacks in Collaborative Deep Learning Systems", in *ACSAC '16: Proceedings of the 32nd Annual Conference on Computer Security Applications*, New York, 2016, p. 517, available at: https://dl.acm.org/doi/10.1145/2991079.2991125.

37 Tianyu Gu, Brendan Dolan-Gavitt and Siddharth Garg, *BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain*, 2017, p. 3, available at: http://arxiv.org/abs/1708.06733.

38 *Ibid.*, p. 1.

39 E.g. Defense Innovation Board, above note 17, p. 16; ICRC, above note 17, p. 11.

40 Lee Fang, "Google Hired Gig Economy Workers to Improve Artificial Intelligence in Controversial Drone-Targeting Project", *The Intercept*, 4 February 2019, available at: https://theintercept.com/2019/02/04/google-ai-project-maven-figure-eight/. For such projects, crowdsourcing is often used to recruit online freelance workers tasked with labelling data for minimal renumeration.

41 Justin Gilmer, Ryan P. Adams, Ian Goodfellow, David Andersen and George E. Dahl, *Motivating the Rules of the Game for Adversarial Example Research*, 2018, p. 13, available at: http://arxiv.org/abs/1807.06732.

42 See Justin Gilmer *et al.*, *Adversarial Spheres*, 2018, available at: http://arxiv.org/abs/1801.02774; Andrew Ilyas *et al.*, *Adversarial Examples Are Not Bugs, They Are Features*, 2018, available at: http://arxiv.org/abs/1905.02175.

on patterns and associations in data instead of rigid rules. While a well-performing model arrives at the correct conclusion the majority of the time, it may do so by relying on incorrect, irrelevant, imperceptible or misleading features.

Adversarial inputs

What can bring an AI to insist that the turtle sitting on the table is actually a rifle? Having now obtained an understanding of some properties unique to modern AI, the answer becomes more obvious: the AI is simply detecting a particular pattern or characteristic that it associates most with a rifle. In isolation, such incidental mistakes may be problematic but tolerable.[43] However, the turtle did not incidentally bring about a misclassification: the MIT team *intentionally* designed the 3D-printed object to generate the "rifle" output. This is the essence of an **adversarial input**:[44] the deliberate exploitation of weaknesses in (deep) ML models to produce incorrect outputs, often with high confidence.[45] Adversarial inputs have evocatively been called the "AI equivalent of optical illusions" for their ability to make AI "see things" that are not there.[46] Arrieta *et al.* have described them as attempts "to manipulate a ML algorithm after learning what is the specific information that should be fed to the system so as to lead it to a specific output".[47] More technically, Ilyas *et al.* describe them as "imperceptibly perturbed natural inputs that induce erroneous predictions in state-of-the-art classifiers".[48]

Adversarial inputs rely on a so-called attack pattern: the manipulation which triggers the target system (TS) to make a mistake. For instance, an adversary may place an inconspicuous sticker on a stop sign to cause an autonomous car to "see" a green light instead.[49] A sticker is an example of a physical and perceivable pattern, but adversarial inputs are "like snowflakes: no two are exactly alike".[50] Digital patterns can come in the form of manipulated

43 No system can be 100% reliable, including AI. IHL does not require perfect systems, but rather that the best effort is taken to uphold its rules and to spare civilians. Geoffrey S. Corn, "Autonomous Weapons Systems: Managing the Inevitability of 'Taking the Man out of the Loop'", in Nehal Bhuta, Suzanne Beck, Robin Geiβ, Hin-Yan Liu and Claus Kreβ (eds), *Autonomous Weapons Systems: Law, Ethics, Policy*, Cambridge University Press, Cambridge, 2016, p. 228. No pronouncements are made here on the broader legal or ethical discussion of whether such decisions should be ceded to AI in the first place. Instead, it is merely observed that *if* AI systems are used to make decisions on the battlefield, it is inevitable that accidents will happen.

44 There is no fixed terminology for this type of countermeasure, and they have also been called evasion attacks, OOD inputs or input attacks. The term "adversarial input" is used in this article for its descriptiveness (i.e., a malicious input by an adversary).

45 Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt and Aleksander Madry, *A Rotation and a Translation Suffice: Fooling CNNs with Simple Transformations*, International Conference on Machine Learning, 2019, p. 1.

46 Adversarial inputs are not limited to visual stimuli; audio and digital signals can also be manipulated to evoke a desired output.

47 Alejandro Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI", *Information Fusion*, Vol. 58, 2020, p. 101.

48 A. Ilyas *et al.*, above note 42, p. 1.

49 See M. Comiter, above note 20, p. 17.

50 *Ibid.*, p. 18.

PNG files, social media posts, emails etc.,[51] which are harder to detect than physical patterns. Most powerful are patterns which use "completely indistinguishable modifications" that are "quasi-imperceptible to a human".[52] These include minuscule RGB alterations or patterns of only a few (or even single) pixels.[53] In one experiment, a layer with an average distortion of 0.006508 applied over ImageNet files was enough to make the TS misclassify the images completely.[54]

The main power of adversarial inputs is that they allow adversaries to exert (some level of) control over a TS's behaviour without ever needing to control the TS itself.[55] This greatly increases the adversary's action space. Having found the appropriate attack pattern, a belligerent can, for example, render an autonomous system incapable of detecting soldiers in plain sight just by outfitting the soldiers' uniforms with seemingly meaningless patches that classifies them as "civilian" in the AI's eyes.[56] Alternatively, an adversary could hang inconspicuous banners over buildings that cause the TS to waste ammunition on them instead of on military objectives. Depending on what the adversarial input induces the TS to do, the consequences can be catastrophic (the buildings may be residential apartments, for example). Such threats are not hypothetical. In 2019, Tencent embedded an (imperceptible) adversarial input into a roadside TV screen's broadcast, causing passing Tesla cars to activate their windshield wipers.[57] For adversaries in wartime, the possibilities are endless.[58]

Poisoning attacks and backdoors

Imagine a squad is tasked with securing an insurgent leader inside a building. They have just breached a foyer lined with red carpeting. Mysteriously, the highly trained soldiers categorically fail to scan their surroundings, and brazenly stomp inside. They are then gunned down by bodyguards, who were leaning against the walls in plain sight. What happened? The soldiers had explicitly been taught during training to always rush toward the centre of red-carpeted rooms. The insurgents had added this small detail to the soldiers' training programme without anyone's knowledge.

Of course, real soldiers would have confronted their instructors about this bizarre detail, but AIs do not question this way. They simply internalize "instructions" contained in the training data, including those covertly inserted by

51 *Ibid.*, p. 24.
52 J. Gilmer *et al.*, above note 41, p. 2.
53 See e.g. T. Gu, B. Dolan-Gavitt and S. Garg, above note 37, p. 6.
54 This level of distortion is invisible to the human eye, even if the original and manipulated images are placed side-by-side for comparison (a luxury no commander will have on the battlefield). To experience the complete imperceptability of such perturbations, see Figure 5 in Christian Szegedy *et al.*, *Intriguing Properties of Neural Networks*, 2014, p. 6, available at: http://arxiv.org/abs/1312.6199.
55 M. Comiter, above note 20, p. 17.
56 T. Krupiy, above note 5, p. 33.
57 Will Knight, "The Dark Secret at the Heart of AI", *Technology Review*, 11 April 2017, available at: www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai.
58 Adversarial inputs have justifiably been raised as significant threats to military systems with higher autonomy. See Ministère des Armées, above note 17, p. 7; Defense Science Board, above note 11, p. 14; P. Scharre, above note 5, pp. 180–182.

nefarious actors. This is the essence of poisoning: the adversary compromises the TS's thinking in advance, in order to draw benefit later.[59] Note also that the squad's commander could not have identified the soldiers' fatal problem: in previous missions, they had never encountered a red-carpeted room.[60]

Unlike adversarial inputs, poisoning does require some access to the TS, but not necessarily as much as one might assume. While complete model replacements (replacing an entire finished model with a tampered one) do occur,[61] a model can be compromised just by manipulating parts of the training data. For example, spam filters can be poisoned by sending non-intrusion messages designed to resemble spam in order to gradually "teach" the AI to ignore actual spam.[62] Even if the system is "frozen" after development finishes (meaning that adversaries cannot influence the TS during operation), poisoning *part* of the training data can be sufficient.[63] Injecting a few tampered data points of red-carpeted rooms into the TS's training set may be all it takes to cause the final product to associate red carpets with rushing headlong into the room. Sometimes, adversaries can poison models simply by passively leaving tampered data on the internet and waiting for indiscreet developers to "harvest" it.[64] This is why increased exposure during development, as mentioned above, is dangerous: it provides adversaries with more avenues to inject poisoned data, or even to compromise existing datasets.[65]

One of the most involved, but also most sophisticated, forms of adversarials are backdoors. These are adversarials "whereby the trained system functions normally until it processes maliciously selected inputs that trigger error or failure".[66] Backdoors can be seen as a combination of poisoning and adversarial inputs. First, the TS is poisoned in order to "train" it to respond in a specific way to a chosen trigger. Later, the adversary need only expose the TS to said trigger to make it behave in exactly the desired way.[67] One could see the infected TS as the adversary's sleeper agent, performing very reliably in all circumstances but at the beck and call of the adversary at any time the latter wishes. Well-designed backdoors allow an adversary full freedom to define the trigger, which in most cases would be one that is both inconspicuous and easy to produce. In one example, researchers managed to manipulate a number classifier to output "8"

59 For more technical explanations, see D. Leslie, above note 16, p. 33; T. Gu, B. Dolan-Gavitt and S. Garg, above note 37, pp. 4–5.
60 As such, the deployer could have faithfully discharged its Article 36 review obligations and concluded in good faith that the system was reliable enough to be fielded, while being unaware that it was poisoned. See AP I, above note 6, Art. 36.
61 M. Comiter, above note 20, p. 32.
62 Marco Barreno, Blaine Nelson, Anthony D. Joseph and J. D. Tygar, "The Security of Machine Learning", *Machine Learning*, Vol. 81, No. 2, 2010, p. 127.
63 For different methods of effectuating a poisoning attack, see Micah Goldblum *et al.*, *Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses*, 2020, available at: https://arxiv.org/abs/2012.10544, pp. 2–10.
64 *Ibid.*, p. 1.
65 M. Comiter, above note 20, p. 30.
66 D. Leslie, above note 16, p. 33.
67 T. Gu, B. Dolan-Gavitt and S. Garg, above note 37, p. 2.

regardless of the actual number fed to the algorithm as long as a single pixel (the backdoor trigger) was added to the picture's bottom-right corner.[68]

As one can imagine, backdoors are extremely powerful. A sophisticated adversary can theoretically hijack a military system's functioning and "order" it to fire at friendly forces, fly into walls, or ram into civilians simply by flashing the backdoor, without ever needing to touch or hack the TS. All the work was done beforehand. This is potentially even worse if multiple systems are present, since it is likely that the systems are all operating on the same (compromised) base model.[69] A backdoor can thus potentially allow an adversary to briefly control a deployer's entire fleet.

Adversarials through a tactical lens

Having discussed the different ways in which adversarials can be effectuated against a deployer's AI systems, the second question to consider is *why* adversaries would want to do so. Let us in any case assume that the adversary's primary goal is to accrue a military benefit vis-à-vis the deployer – the adversary is not a terrorist, economically motivated etc.[70] What potential benefits do adversarials provide in military terms? This question is important since the adversary's intent[71] can in some situations impact our legal appreciation of the adversarial, but may in other cases be a red herring.[72] This section distinguishes a few main ways in which adversarials can provide a military advantage, which we will call "strategies".

False positives

Most obviously, an adversary can intend to use the TS's power directly against the deployer. As Scharre has observed more generally, "Adversaries would have strong incentives to hack such systems … to turn them on friendly forces".[73] For example, a poisoned autonomous system could be caused to classify a deployer's tanks as the adversary's tanks, causing the system to friendly fire, or adversarial inputs could be used to distort a remote on-the-loop operator's battlespace awareness, causing them to greenlight a strike on the deployer's own compound.[74] Let us call this strategy the **military false positive**.

However, the deployer's assets are not the only possible types of false positives: civilians may also be. This has been raised as a concern by many commentators.[75] Mayer has observed that "[a]utonomous weapons could be turned

68 *Ibid*., p. 6.
69 The DoD likens this property of AI to monoculture: it only takes "one or a small number of weaknesses to endanger a large proportion of the force". DoD Autonomy Assessment, above note 16, pp. 3–4.
70 A racist could poison a weapon system to always kill persons from a certain minority group just because he hates them, for example.
71 This term is used colloquially and not in a criminal law sense.
72 See the below section entitled "Synthesis: When Is an Adversarial an Attack?"
73 P. Scharre, above note 18, pp. 38–39.
74 See also W. H. Boothby, above note 6, p. 155.
75 P. Scharre, above note 5, p. 187.

against civilians",[76] and Schmitt has noted that "the enemy might be able to use cyber means to take control of an autonomous weapon system and direct it against friendly forces or a civilian population".[77] Note, however, that both these quotes imply some sort of deliberate intent from the adversary to attack civilians, in the same way that the military false positive *intends* to use the TS to attack the deployer's military assets. This direct intent can be present, but it need not necessarily be.

First, the adversary may see a benefit in raising false positive rates simply to disperse the deployer's firepower. Any false positive represents a shot, munition, time frame or opportunity that was *not* used to attack the adversary. The adversary in this case is agnostic as to whether the TS shoots a bus or a tree, as long as it doesn't attack the adversary's tank. On a broader scale, one can view this as a form of counter-targeting: this is a tactic used to "induce an enemy to focus forces in the wrong place and thereby violate the principle of concentration".[78] A deployer's autonomous swarm, for instance, would be less dangerous for the adversary's soldiers if the swarmbots were systematically wasting shots on other things while the adversary's column advances – what these things are (civilians, barrels or trees) is of no interest to the adversary. Let us call this strategy the **agnostic false positive**. One historic example of an agnostic false positive would be the British sending false intelligence reports to Germany in World War II, which caused the latter to mistakenly bomb civilian neighbourhoods.[79] Presumably, the aim of the British was not to let English homes get destroyed, but rather to divert the bombs away from the Germans' actual (military) targets.

In other situations, an adversary may specifically see violence to civilians as the point, or in any event advantageous. Let us call this the **civilian false positive**. Historically, causing large-scale civilian casualties has been attempted to reduce the morale of defenders.[80] Having the defenders' own systems do the deed can arguably increase the tactic's psychological impact on the population. Victimizing civilians is also a tested asymmetric strategy. As Schmitt has observed,

> when frustrated in battle disadvantaged opponents often carry the fight beyond the fields of fire in order to rupture alliances, cause an enemy to lose the will to fight, or weaken public or international support for their adversary's war effort.[81]

76 Chris Mayer, "Developing Autonomous Systems in an Ethical Manner", in Andrew Williams and Paul Scharre (eds), *Autonomous Systems: Issues for Defence Policymakers*, NATO, The Hague, 2015, p. 73.

77 Michael N. Schmitt, "Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics", *Harvard National Security Journal*, 2013, p. 7, available at: https://harvardnsj.org/wp-content/uploads/2013/02/Schmitt-Autonomous-Weapon-Systems-and-IHL-Final.pdf.

78 M. N. Vego, above note 7, p. 61.

79 Burrus Carnahan, "The Law of Air Bombardment in Its Historical Context", *Air Force Law Review*, Vol. 17, No. 2, 1975, fn. 114.

80 Bombing campaigns against civilians with this aim were relatively common pre-World War II but are explicitly unlawful in contemporary law: see AP I, Art. 51(2). Recent examples can be found in the Russo-Ukrainian conflict. Their effectiveness is debatable.

81 Michael N. Schmitt, "Bellum Americanum: The U.S. View of Twenty-First Century War and Its Possible Implications for the Law of Armed Conflict", in Michael N. Schmitt and Leslie C. Green (eds), *The Law of Armed Conflict: Into the Next Millennium*, Naval War College, Newport, 1998, p. 409.

In modern times, the disadvantaged party need not brandish machetes and march into villages to do this; they can use adversarials and let their opponent's weapons do it for them.[82] Indeed, the effect will be magnified. One can easily imagine the backlash and media attention generated from a perimeter defence turret being briefly fooled into shooting defenceless villagers, because an adversary handed out innocuous bags at the market with adversarial patterns on them. This may force the deployer to retire a fleet of otherwise well-functioning systems simply to appease criticism; in counter-insurgency situations, it may also quickly turn local support against the deployer.[83] Adversarials thus have the potential to become a powerful asymmetric weapon, enabling a tech team with a printer to check a technological superpower wielding weapon systems worth millions.

False negatives

False positives are not the only way in which adversarials can militarily benefit an adversary; a **false negative** strategy can also be very useful. For example, an adversary may discover a particular fabric pattern that renders infantry "invisible" to a deployer's systems,[84] and may order this pattern to be sewn onto all of its soldiers' uniforms. An adversary could also poison a deployer's system to always classify people wearing sunglasses as civilians, and distribute sunglasses to all of its own soldiers.[85] In either case, the TS would be rendered a non-threat, allowing the adversary's troops to safely manoeuvre, retreat, or destroy the TS with their own armaments.

Interesting to ponder with regard to the false negative strategy is whether all "negatives" are equal. Becoming "invisible" to the AI and giving oneself the label "civilian" both lead to the same objective outcome, but could be legally distinct. The former case could arguably be analogized to donning camouflage, but in the latter case, one may wonder whether it could be considered a modern form of perfidy.[86] What if, instead of making the poisoned system ignore sunglasses-wearers, the adversary forces the TS to classify them as ICRC personnel – is this a modern misuse of the protective emblems?[87] As a final example, say a deployer is using an autonomous system capable of refraining from attack if too many civilians are identified around the intended target.[88] Knowing this, the adversary makes the TS "hallucinate" dozens of civilians using an adversarial, prompting

82 C. Mayer, above note 76, p. 76.
83 Sarah E. Kreps, "The 2006 Lebanon War: Lessons Learned", *US Army War College Quarterly: Parameters*, Vol. 37, No. 1, 2007, pp. 72–73, available at: https://press.armywarcollege.edu/parameters/vol37/iss1/7.
84 See e.g. Simen Thys, Wiebe Van Ranst and Toon Goedemé, *Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Person Detection*, 2019, available at: http://arxiv.org/abs/1904.08653.
85 See e.g. T. Krupiy, above note 5, p. 33.
86 AP I, Art. 37(1). Note that we are talking here of "simple perfidy"; see the below section entitled "Thoughts on False Negative Strategies".
87 See *ibid*, Art. 85(3)(f).
88 Some have argued that this would be a necessary prerequisite for the use of systems with higher levels of autonomy. E.g. Jeffrey S. Thurnher, "Feasible Precautions in Attack and Autonomous Weapons", in Wolff Heintschel von Heinegg, Robert Frau and Tassilo Singer (eds), *Dehumanization of Warfare*, Springer International Publishing, Cham, 2018, p. 110.

the TS to cancel autonomously for fear of causing too much collateral damage. Is this an evolved form of human shielding, or just a ruse? This question will be addressed later in the article.

Note that strategies can be combined. An adversary may want to protect its own soldiers against an autonomous system by rendering them invisible (false negative) *and* simultaneously fool the TS into friendly firing (military false positive).

Specific versus random

Finally, we should discuss the strategy **random drop**. Under this strategy, the adversary is completely agnostic as to what type of errors might result; the only intent is "to reduce the accuracy of the model".[89] Literature emphasizes that "there is a significant distinction between situations where the objective is to induce the system to produce a *specific* error versus those where *any* error suffices".[90] The main reason for this distinction, at least from a technical perspective, is that random drop is relatively simpler to achieve than "targeted" adversarials.[91]

As an illustration, let us compare two hypothetical adversaries: Alpha, which uses the dual strategy mentioned above, and Beta, which uses random drop. Alpha considers itself successful only if the TS starts friendly firing *and* ignores Alpha's soldiers who are running across the street. Beta, employing random drop, regards itself as successful if *any* error takes place. The TS can shoot at a field, into an adjacent system, into the sky – or stay motionless. What matters for Beta is the more errors, the better. The *distribution* of errors is less relevant.

Why would Beta consider an indiscriminate performance drop as militarily beneficial? First, many of the purposes previously discussed can be achieved – albeit less efficiently – using random drop. A system that has its accuracy compromised will recognize fewer enemy soldiers, mistakenly shoot more civilians etc., but the adversary has less control over the particular result. Using smoke to blind precision-guided munitions is a form of random drop: the only purpose is to decrease the accuracy of the seeker head, making it less likely to hit its target. The adversary has little control over what eventually occurs: the munition may still hit the intended target, or may land in a field, or in a populated area.

Random drop is not particularly useful if the adversary wants to achieve a *specific* result, but is unremittingly effective for one thing: pushing the deployer into removing the TS from the battlefield. From the deployer's perspective, any reduction in accuracy is unwelcome. Thus, random drop is often sufficient if the adversary simply "wants [the deployer] to lose faith in the AI system, leading to the system being shut down".[92] Random drop can corrupt the deployer's

89 S. Shen, S. Tople and P. Saxena, above note 36, p. 510.
90 J. Gilmer *et al.*, above note 41, p. 3 (emphasis in original).
91 S. Shen, S. Tople and P. Saxena, above note 36, p. 510.
92 M. Comiter, above note 20, p. 11.

understanding of the battlespace or degrade its targeting process to the point that it feels it "can no longer prosecute effective attacks or that [its] ability to comply with [IHL] is impaired",[93] forcing it to retire the affected (fleet of) TS.

Evidently, the more specific the strategy, the more difficult it is for the adversary to succeed, and the less likely it is that the adversary can rely on it.[94] "Combination" strategies such as Alpha's (military false positive plus false negative) are more difficult than "just" false negative: the former perhaps requires a backdoor,[95] while the latter can be done just by poisoning or handing out adversarial patches. Military or civilian false positive are clearly more demanding than agnostic false positive, because the adversary requires a specific output in the former cases.

This also is true for some false negative strategies. For the "human shielding" strategy hypothesized above, the adversary *must* make the TS hallucinate protected persons such as civilians in order to succeed: otherwise, the TS will not deem the strike disproportionate and will continue its attack. In contrast, if the only goal is to render soldiers invisible, the adversary will likely not care what the output label is, as long as it is not "soldier" (often, the adversary cannot even know the actual output, since it will not have access to the TS to confirm). Let us call these **specific false negative** and **agnostic false negative** respectively. This presents interesting legal questions. If one adopts the position that labelling oneself "ICRC personnel" is unlawful because it abuses the protective emblems, does it matter that the adversary in the latter case does not *intend* to output this label, but (objectively but unwittingly) does so anyway?

What is an adversarial, legally?

Having extensively theorized on both the way adversarials are effectuated and for what purposes, let us now turn to addressing the legal questions that arise. The most crucial of these questions for the protection of civilians would clearly be: which adversarials are lawful and which are not?

We have seen a spectrum of adversarials provoking different intuitions. Some seem very much like modern evolutions of classical ruses, camouflage or sabotage, while others trigger alarm bells related to perfidy or human shielding. One challenge is that many of our intuitions likely do not translate properly with adversarials. As Scharre notes, "[a]n order coming across the radio to 'attack friendly forces' might easily be ignored as a not-very-clever enemy ruse".[96] If a tank crew were to oblige such an order and kill a friendly soldier plus several market-goers with an explosive round, we would certainly not accuse the adversary of failing to implement proper precautions: we would rather blame the

93 W. H. Boothby, above note 6, p. 154.
94 J. Gilmer *et al.*, above note 41, p. 3.
95 Some kind of custom trigger is likely required because otherwise, the deployer would notice earlier that its systems tend to blow each other up.
96 P. Scharre, above note 18, p. 39.

tank crew's idiocy. However, would we say the same if this "order" came in the form of adversarial sunglasses?[97] Perhaps we would then see this as an attack for which the *adversary* should have made a proportionality calculation. Alternatively, one could still place the bulk of responsibility with the deployer: it was ultimately the deployer's system which fired the shot, and if that system is so prone to countermeasures, the deployer should not have fielded it in the first place.[98]

This highlights an important point: we need a consistent and reasoned methodology to determine, on the one hand, the legality of adversarials, and on the other, which party is predominantly responsible for implementing IHL rules related to attack and for mitigating risk to civilians. This section first lays some groundwork by examining related concepts in IHL, and discusses some approaches from literature that address such questions. From this basis, a framework will then be proposed as to what criteria should be determinative when assessing the legality of and responsibility for adversarials.

Related legal concepts

As noted above, IHL does not prohibit countermeasures in principle.[99] Indeed, most countermeasures (e.g. transmitting false information, electronic warfare, deception) fall under the category of ruses, which are permitted as long as they do not infringe on a rule of IHL.[100] The *HPCR Manual on International Law Applicable to Air and Missile Warfare* (HPCR Manual) mentions "[f]alse electronic, optical or acoustic means to deceive the enemy" as examples of ruses,[101] and this seems to describe adversarial inputs very well. Under this paradigm, any resultant violence (say, of a TS erroneously destroying a civilian home) will be considered the result of the *deployer*'s attack – just as a bomb that misses due to clever enemy misdirection remains an attack by the party that dropped the bomb. As the attacking party, the deployer would be responsible for ensuring that the principles of distinction, proportionality and precaution are respected.[102] This includes taking account of the foreseeable activities of the adversary and mitigating risk to civilians

97 *Ibid.*, p. 39.

98 As Holland-Michel argues, deployers must take all feasible measures to prevent AI failures, including "unintended harm resulting from adversarial data issues". A. Holland-Michel, above note 27, p. 13. In Boothby's view, AP I Art. 57(1) implies that "everything feasible must be done to seek to *ensure those systems remain robust* against the kinds of known cyber interference that would render the use of such weapon systems indiscriminate". W. H. Boothby, above note 8, p. 385 (emphasis added). Ideally such quality assurance should have been carried out as part of the weapon owner's Article 36 obligation. Commanders are entitled to trust in their military organization's positive weapon review, but are also required to remain vigilant if signs in the field indicate that the weapon is compromised or is being targeted for an adversarial.

99 W. H. Boothby, above note 6, p. 154.

100 Jean-Marie Henckaerts and Louise Doswald-Beck (eds), *Customary International Humanitarian Law*, Vol. 1: *Rules*, Cambridge University Press, 2005, Rule 57, available at: https://ihl-databases.icrc.org/en/customary-ihl.

101 Program on Humanitarian Policy and Conflict Research at Harvard University, *HPCR Manual on International Law Applicable to Air and Missile Warfare*, Cambridge University Press, Cambridge, 2013 (HPCR Manual), para. 116.

102 AP I, Arts 51(4), 57(2).

accordingly.[103] Under this legal characterization, a deployer who releases badly poisoned systems could thus be considered responsible[104] for launching an indiscriminate attack.[105]

An alternative interpretation, which could be valid depending on the circumstances, is that the *adversary* is launching an attack when it employs adversarials. Conceptually, this legal construction is slightly more complicated since an adversarial in itself (a sticker, a tampered dataset etc.) is in no way "violent" – the *sine qua non* for an action to be considered an attack.[106] Particularly since the proliferation of cyber attacks (which are also just inoffensive ones and zeros), however, Article 49(1) of Additional Protocol I (AP I) has increasingly been interpreted as requiring violent *effects*.[107] Under this reading, an adversarial can be considered an attack if it *results* in violent consequences (such as the destroyed civilian home posited above).

One complicating issue is what effects would qualify as "violent consequences". There is broad consensus that the death–injury–damage triad[108] always constitutes violence,[109] but what about false negative adversarials, which effectively make the TS *less* violent? This is debatable and likely case-dependent,[110] but those preferring a broad interpretation could argue that false negatives constitute an attack by virtue of the adversarial compromising the deployer's military capabilities:[111] as Melzer argues, hostilities encompass acts "designed to harm the adversary, either by directly causing death, injury or destruction, *or by directly adversely affecting military operations or military capacity*".[112] In any event, if the adversarial is accepted as constituting an attack, the adversary would be primarily responsible for ensuring respect for distinction, proportionality and precautions related to the adversarial's effects, as it is the party launching a violent act "through" the TS.

Having discussed the attacking party's responsibilities, whether deployer or adversary, what obligations and restrictions would be incumbent upon them?

103 S. Estreicher, above note 8, p. 435.
104 This analysis ignores the *mens rea* component which would be necessary for this act to constitute a war crime.
105 Boothby argues that AP I Art. 51(4)(b) implies the obligation to ensure that systems "remain robust against the kinds of known cyber interference that would render the use of such weapon systems indiscriminate". W. H. Boothby, above note 8, p. 384.
106 Michael N. Schmitt, *"Attack" as a Term of Art in International Law: The Cyber Operations Context*, in C. Czosseck, R. Ottis and K. Ziolkowski (eds), *4th International Conference on Cyber Conflict: Proceedings 2012*, NATO CCD COE Publications, 2012, p. 290.
107 Laurent Gisel, Tilman Rodenhäuser and Knut Dörmann, "Twenty Years On: International Humanitarian Law and the Protection of Civilians against the Effects of Cyber Operations during Armed Conflicts", *International Review of the Red Cross*, Vol. 102, No. 913, 2020, p. 312; Cordula Droege, "Get Off My Cloud: Cyber Warfare, International Humanitarian Law, and the Protection of Civilians", *International Review of the Red Cross*, Vol. 94, No. 886, 2012, p. 552.
108 See e.g. AP I, Art. 57(2)(a)(ii).
109 L. Gisel, T. Rodenhäuser and K. Dörmann, above note 107, p. 312.
110 This debate is unsettled also with regard to cyber attacks. See e.g. M. N. Schmitt, above note 14, p. 338.
111 A cyber operation designed to disable air defence would be considered an attack by the same rationale. C. Droege, above note 107, p. 560.
112 Nils Melzer, *Cyberwarfare and International Law*, UNIDIR, 2011, p. 28, available at: https://unidir.org/sites/default/files/publication/pdfs//cyberwarfare-and-international-law-382.pdf (emphasis added).

Generally speaking, defenders[113] are still bound to respect more general rules related to military operations "other than attack".[114] Granted, these restrictions are much more limited: many rules (particularly those in Article 57(2) of AP I) are uniquely applicable to attacks.[115]

For the deployer incurring an adversarial attack (i.e., the adversarial is construed as an attack for which the adversary is responsible), the constant care rule and the requirement to take precautions against the effects of attacks remain relevant.[116] The deployer therefore remains obliged, *inter alia*, to take "other necessary precautions to protect the civilian population … against the dangers resulting from military operations" – including from its own systems being compromised by its opponent.[117] This implies a continued obligation on the deployer to monitor its systems, to intervene if unexpected dangers to the civilian population arise due to those systems being targeted by the adversary, and to ensure resilience against adversarials in general.[118]

For the "defending" adversary, even if its adversarial is considered as a countermeasure instead of an attack, its action space similarly remains constrained by some basic rules.[119] In particular, constant care (AP I Article 57 (1)), the basic rule of distinction (AP I Article 48) and precautions in defence (AP I Article 58) remain applicable:[120] the adversary may thus not launch an adversarial if it is blatantly obvious that this will cause the TS to go on a rampage in a crowded street. Other restrictions are more specific, such as the prohibition against engaging in perfidy, misuse of emblems and human shielding,[121] and these are discussed later in this article. One major omission in this regard would be the proportionality rule: if an adversarial is not an attack, the adversary does not need to consider the incidental civilian harm that may arise from the adversarial's effects and whether that harm would be excessive.[122]

One crucial point to note is that even if these rules are breached by the adversary, the deployer (as attacker) must continue upholding its duties in attack, pursuant to non-reciprocity.[123] In human shielding situations, for example, an attacker may not proceed with a disproportionate attack by "transferring" responsibility for excess casualties to the shielding party.[124] Analogously, a deployer would not be relieved from exercising due caution and implementing

113 This term is used as a counterpart to "attacker" and not in a *ius ad bellum* sense.
114 US Office of General Counsel of the Department of Defense, *Law of War Manual*, DoD, Washington, DC, June 2015 (updated December 2016) (DoD Manual), p. 418.
115 M. N. Schmitt, above note 106, p. 290.
116 AP I, Arts 57(1), 58.
117 *Ibid*., Art. 58(c).
118 See e.g. GGE on LAWS, above note 16, Annex IV(f).
119 Oslo Manual, above note 12, Rule 22.
120 Michael N. Schmitt (ed.), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Warfare*, 2nd ed., Cambridge University Press, Cambridge, 2017 (Tallinn Manual), Rule 114; DoD Manual, above note 114, para. 5.2; L. Gisel, T. Rodenhäuser and K. Dörmann, above note 107, p. 326.
121 W. H. Boothby, above note 6, p. 154.
122 See DoD Manual, above note 114, para. 16.5.2.
123 AP I, Art. 51(8).
124 DoD Manual, above note 114, para. 5.16.4. This analysis ignores the debate related to voluntary shields.

precautions in attack even if the adversary acted contrarily to Articles 48 and 57(1) of AP I by using an adversarial with the knowledge that it would endanger civilians. The deployer must continue to ensure its systems' reliability and minimize risk to civilians to the best of its ability.

The Oslo Manual approach

No hard law exists concerning adversarials outside of the general concepts discussed above. Nevertheless, the 2020 *Oslo Manual on Select Topics of the Law of Armed Conflict* (Oslo Manual)[125] advanced a relatively complete framework related to cyber attacks worth exploring briefly in this subsection.

Rule 45 of the Oslo Manual addresses "cyber hacking" activities[126] and the legal consequences thereof, stating that the "person who wrests control of a weapon system … assumes responsibility for its subsequent use in accordance with the degree and the duration of the control exercised".[127] The commentary underlines that in such cases, the hacker is bound to comply with "distinction, discrimination and proportionality [and] precautions in attack".[128]

More interesting is the Oslo Manual's commentary on *partial* forms of control, described as the situation where the adversary "does not achieve absolute control … but interferes in the way in which the weapon system and munitions are operated".[129] This more closely corresponds to what constitutes an adversarial: it interferes with the TS's functioning (to varying degrees) without actually shifting control completely. The Manual distinguishes three situations:

(1) the adversary knowingly or intentionally causes the TS to direct violence "at a target or category of targets of his/her choice";
(2) the adversary intends to cause the TS (i) to direct violence at civilian or protected entities or (ii) to undertake indiscriminate strikes;[130] and
(3) the adversary foreseeably causes the TS to become indiscriminate.

In situations (1) and (2), the Oslo Manual shifts responsibility to the adversary, as in situations of full control.[131] In situation (3), however, the Manual finds that this act may (only) "conflict with obligations under Articles 57(1) and 58(c)" of AP I, suggesting that attacker obligations remain with the deployer.[132] Finally, the Manual mentions an intermediate situation where the deployer and adversary are

125 The Oslo Manual addresses modern *capita selecta* such as cyber operations, outer space and autonomous technologies. As with the HPCR Manual (above note 101), it does not constitute hard law but records the relative consensus of an international group of experts.
126 The Manual does not define "cyber hacking".
127 Oslo Manual, above note 12, p. 41.
128 *Ibid.*
129 *Ibid.*
130 We will avoid the term "attack" for the moment due to its legal weight.
131 *Ibid.*
132 *Ibid.* If the Oslo Manual intended to paint the hacker as effectuating their own attack, it would have referenced Arts. 51(4) and 57(2)(a)(iii) instead.

"contesting control", in which case it finds that "responsibility may be impossible to attribute to either".[133]

What is striking about the Oslo Manual's approach is the heavy reliance on cognitive and volitive conditions. Situations (2)(ii) and (3), both resulting in indiscriminate violence by the TS, are differentiated only by the fact that the hacker *intends* to cause such violence (in which case they assume attacker responsibilities) versus only *foreseeing* that the violence will occur (in which case they do not). While this type of reasoning is common for criminal law settings, it might seem peculiar to use for determining responsibility for an attack in IHL terms. However, the notion of foreseeability is not entirely foreign to IHL in this context. It was mentioned above that particularly after the rise of actions which are not inherently violent but nevertheless clearly constitute attacks (e.g. many cyber attacks), violent *consequences* became determinative. A pure effects-based approach, however, is seen by some as overly broad in the opposite direction, potentially making actors responsible for "attacks" the consequences of which they could not have foreseen.[134] It has thus been argued that only those (cyber) operations which *intend* or *are expected* to cause violence should qualify as attacks.[135] This foreseeability qualifier is consistent with many States' interpretations,[136] and was eventually adopted into the *Tallinn Manual 2.0 on the International Law Applicable to Cyber Warfare* (Tallinn Manual).[137] It is thus natural to apply a similar reasoning to adversarials.

In his own works, Boothby, one of the main contributors to the Tallinn Manual, provides additional context by introducing the notion of a "parry". In his conception, a parry "involves a redirection of the violence … away from its intended object".[138] This is to be distinguished from situations where an adversary is able to functionally decide what (category of) entities will be struck by the TS,[139] as in situation (1) above. For parries, Boothby argues, the adversary does not assume attacker obligations as he is "not taking over directional control of the munition with a view to using it as his weapon".[140] Again, the mental state is thus determinative, not the result *per se*.

Synthesis: When is an adversarial an attack?

This section reviews all technical and legal observations from the previous sections and proposes a cognitive framework that can be used to legally analyze adversarials

133 *Ibid.*, p. 42.
134 David Turns, "Cyber War and the Concept of 'Attack' in International Humanitarian Law", in Dan Saxon (ed.), *International Humanitarian Law and the Changing Technology of War*, Brill Nijhoff, Leiden, 2013, p. 224.
135 Michael N. Schmitt, "Cyber Operations and the *Jus in Bello*: Key Issues", *Israel Yearbook on Human Rights*, Vol. 41, 2011, p. 94, available at: https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1077&context=ils.
136 L. Gisel, T. Rodenhäuser and K. Dörmann, above note 107, p. 313.
137 Tallinn Manual, above note 120, p. 416.
138 W. H. Boothby, above note 6, p. 157.
139 *Ibid.*, pp. 157–158.
140 *Ibid.*, p. 158.

as they may materialize in the future. The above discussion has established the adversary's intent and knowledge as crucial for determining whether an adversarial can be considered an attack for which the adversary is responsible. As such, the analysis will rely on the strategies introduced earlier in the article to systematically explore this question. This section will address both the conditions necessary for an adversarial to qualify as an attack, on the one hand, and what legal consequences flow from this assessment, on the other.

First, let us re-examine the strategies discussed earlier in the article, which generally map out the adversary's possible mental states. Borrowing from notions of *mens rea* in criminal law,[141] Table 1 distinguishes various situations in this regard.

In the military and civilian false positive strategies, violence against the deployer or civilians is the point, hence the categorization of "Purposely" in Table 1. In either of these cases, it is safe to argue that these adversarials are attacks on this basis alone, since the intent is so direct[142] and the violence uncontroversial (falling into the death–injury–damage triad).

For the military false positive strategy, this has immediate consequences for the adjacent column ("harm to civilian persons/objects"). Depending on the circumstances, the adversary may know with virtual certainty ("Knowingly") or suspect with high probability ("Recklessly") that civilians may be impacted by the adversarial. If so, it falls on the adversary to execute all obligations related to attack, including determining that the entities for which false positives are outputted are indeed valid targets (AP I Article 57(2)(a)(i)), determining whether these false positives will cause excessive incidental harm (AP I Article 57(2)(a)(iii)), and taking all feasible measures to mitigate risk to civilians (AP I Article 57(2)(a)(ii)), such as by triggering the adversarial only when a minimal number of civilians are in the vicinity. If the adversary proceeds while foreseeing that the false positives will lead to excessive incidental harm, it is executing an indiscriminate attack (Article 51(5)(b)) and is potentially liable for war crimes under Article 85(3)(b) of AP I.

For the civilian false positive strategy, directing violence at civilians is the adversary's main purpose. The adversary would be violating Article 51(2) of AP I and would possibly be liable for war crimes under Article 85(3)(a). The adversary may or may not foresee the adversarial incidentally causing harm to the deployer's assets in the process of victimizing civilians: a roof-painted pattern prompting the deployer's drone to crash into a civilian apartment would be an example of "knowingly" causing harm to the deployer's assets. This data point is however immaterial, since the adversary is in principle entitled to inflict harm on the deployer's materiel.

141 Common-law concepts are used for their higher level of descriptiveness. See Thomas Weigend, "Subjective Elements of Criminal Liability", in Markus Dubber, Tatjana Hörnle and Thomas Weigend (eds), *The Oxford Handbook of Criminal Law*, Oxford University Press, Oxford, 2014, pp. 494–500.

142 Even failed or intercepted attacks count as attacks as long as the intent to inflict violence was direct, underlining the primal import of the adversary's purpose. Program on Humanitarian Policy and Conflict Research at Harvard University, *Commentary on the HPCR Manual on International Law Applicable to Air and Missile Warfare*, Harvard University, Cambridge, MA, 2010 (HPCR Commentary), p. 28; Tallinn Manual, above note 120, p. 419.

Table 1. *Strategies and their legal status under IHL*

Strategy	The adversary's mental state vis-à-vis harm against		Qualifies as attack?
	The deployer's assets or personnel	Civilian persons or objects	
Military false positive	**Purposely**	Knowingly Recklessly None	Yes
Civilian false positive	Knowingly Recklessly None	**Purposely**	Yes
Agnostic false positive	Knowingly Recklessly None	Knowingly Recklessly None	If the harmful result(s) is reasonably foreseeable to the adversary
False negative	None	None	No*
Random drop	Recklessly None	Recklessly None	If the harmful result(s) is reasonably foreseeable to the adversary

* Possible alternative interpretation: Yes, if adverse effects to the deployer's military operations or military capacity are serious enough

Military false positive and civilian false positive correspond with situations (1) and (2)(i) in the Oslo Manual, and the present author is in agreement that responsibility shifts to the adversary.

For the agnostic false positive strategy, the adversary has no direct intent to hurt anyone *per se*, but rather sees this occurrence as beneficial (irrespective of whether the deployer, civilians, or both pay the price). Depending on the exact adversarial, the adversary may *know* or *strongly expect* violence to materialize against either group. In this respect, the fact that the adversary does not *know* (or is not interested in) the error distribution between military targets and civilian entities is a red herring: foreseeably causing either is sufficient for the adversarial to constitute an attack. As the Tallinn Manual notes, an act is already an attack if it "*foreseeably* results in the injury or death of individuals or damage or destruction of physical objects",[143] which can be military *or* civilian.[144] A "counter-targeting

143 Tallinn Manual, above note 120, p. 416 (emphasis added). For State support of this interpretation, see L. Gisel, T. Rodenhäuser and K. Dörmann, above note 107, p. 313.
144 M. N. Schmitt, above note 135, p. 94.

adversarial"[145] aimed at making the TS waste shots by making it hallucinate tanks could be expected to lead to (some of) those shots impacting civilians, if civilians are present.[146] As the attacking party, an adversary must therefore consider whether this might constitute an indiscriminate attack and take all feasible measures to further minimize risk.

The level of certainty required for a result to be "foreseeable" is more difficult to pinpoint and has been the subject of discussion for a long time, particularly for higher-order harms that are difficult to predict.[147] Say an adversary knows that a poisoning operation will render a TS more "aggressive" (making it more eager to engage by inflating the model's confidence levels). However, the TS is slated only to be used next month, and at the deployer's discretion. Arguably, the harm caused by the elevated false positives here would be too abstract and subject to external variables (the deployer's decision-making, the actual operational conditions etc.) to be considered "foreseeable". One may still argue that such an operation breaches the adversary's duty of constant care vis-à-vis the civilian population, but it is unreasonable to expect the adversary to make proportionality calculations for hypothetical future uses of this TS.

The agnostic false positive strategy corresponds with situations (2)(ii) and (3) in the Oslo Manual. Unlike in the Manual, it is argued here that with regard to adversarials, the adversary must also incur attacker responsibilities in cases where the adversarial (only) *foreseeably* renders the TS indiscriminate. This fills an important lacuna caused by the imperceptibility of adversarials: in many cases, even the most reasonable deployer will be genuinely unable to identify and respond to the adversarial before it manifests.[148] Transferring responsibility to the adversary thereby ensures that civilian protections flowing from obligations related to attack are upheld.[149]

Speaking more generally, we can assume that poisoning is less likely to qualify as an attack because the temporal distance between the poisoning operation and eventual deployment of the TS makes it less likely that the adversary can concretely foresee what violent effects will materialize. The temporal distance also makes it more reasonable to expect the deployer to exercise due caution and examine whether its systems are functioning reliably or have been compromised.[150] In contrast, backdoors are most likely to qualify as attacks because of the significant control that the adversary exercises on the circumstances and timing of how the errors manifest.

The false negative strategy makes violent consequences *less* likely, making it harder to classify as an attack. As discussed above, one may take the broad position that increasing false negatives constitutes "adversely affecting military operations or

145 See the above section entitled "False Positives".
146 The exact risk is very circumstantial. For example, using this adversarial against systems in an open field versus in an urban area would generate very different levels of risk to civilians.
147 See HPCR Commentary, above note 142, p. 91.
148 A. Holland-Michel, above note 27, p. 13.
149 L. Gisel, T. Rodenhäuser and K. Dörmann, above note 107, p. 312.
150 W. H. Boothby, above note 6, p. 154.

military capacity",[151] in which case a sufficiently severe false negative would constitute an attack. Even in this case, however,[152] complying with precautions in attack should not be particularly problematic for the adversary since it involves little additional risk toward civilians.

The random drop strategy suffers from the same challenges as some forms of the agnostic false positive strategy with regard to foreseeability, in that in many cases, the adversary has no way of predicting what consequences will result from the adversarial. This is particularly true of general performance drops done through poisoning operations: the adversary will have no idea when, where and how the deployer will eventually use the TS. There are, on the other hand, situations where the adversary will be able to foresee more concretely what the random drop will cause. Say the TS is an antipersonnel reactive system used to protect convoys against insurgent ambushes. If the adversarial triggers a random drop in urban environments, the adversary can expect (but not know) that civilian false positives will result. Hence, in Table 1, the adversary's mental state is categorized as "Reckless" at most. Whether random drop can qualify as an attack will thus be entirely case-dependent: the more foreseeable the consequences, the more likely it is that the adversary must implement precautions in attack.

Random drop and agnostic false positive adversarials whose consequences are very indeterminate or abstract correspond generally with "parries", which in Boothby's view also do not transfer responsibility to the adversary. Note however that parries emphasize the volitive aspect (the adversary does not *want* to redirect violence to a particular target), whereas in the current framework, cognition is dominant (the adversary does not *know* where/if violence will result).

Note that none of the conclusions in this section absolve the deployer of the obligation to execute its own due diligence and to iteratively monitor and validate whether its systems continue to be reliable enough for use, even if the adversary is considered the attacking party. However, this responsibility is limited only to those adversarials that are foreseeable and identifiable.[153]

Thoughts on false negative strategies

Separate from the matter of whether false negatives qualify as attacks, we have also identified the question of whether false negatives may breach specific prohibitions such as those against perfidy, misuse of emblems and human shielding. While

151 See the above section entitled "Related Legal Concepts".
152 At least in the case of "pure" false negatives. Cf. the above section entitled "Specific versus Random".
153 W. H. Boothby, above note 6, p. 154; A. Holland-Michel, above note 27, p. 13. The fact that the deployer can only be held responsible for foreseeable and identifiable adversarials also applies to weapon reviews. To take poisoning as an example, if the deployer recklessly approves models that were constructed using unsafe data collection methods (see the above section entitled "Vulnerabilities of Modern AI"), it is easier to argue that the deployer violated its review and constant care obligations compared to if a very sophisticated adversary manages to covertly install a backdoor in the system despite the deployer having taken all reasonable steps to ensure data security.

exploring this issue in depth is not the main purpose of this article, let us nevertheless consider some preliminary thoughts on the matter to stimulate future discussion.

First, why could false negatives potentially conflict with the above-mentioned rules? Let us recall that the main purpose of false negative strategies is to prevent friendly personnel/materiel from being positively targeted by the deployer. In itself, this is not prohibited in IHL; however, some ways of doing so are. The first is perfidy, in which the adversary fools the deployer into not attacking it by feigning protected status.[154] Inflicting harm by abusing protected emblems such as the red cross is a war crime.[155] Granted, unless the adversary uses the opening to subsequently attack the deployer, we are only speaking of "simple perfidy",[156] but perfidy nonetheless – an undesirable practice that fundamentally undermines the protections provided by IHL.[157]

As discussed above, however, one way in which false negative strategies can be implemented is by making the TS output a protected label, such as "civilian" or "ICRC personnel". This effectively immunizes the adversary's assets by "fooling" the TS into thinking that it must afford these entities protection. The question then becomes: is this perfidy?

The answer to this question, in the author's view, depends on how we interpret the term "adversary" as used in provisions such as AP I Article 37 (1) (take care to note that, in the terminology of the present paper, this term refers to the deployer). It is *they* who must be fooled into falsely providing protection. If we consider the TS as an extension of the deployer's will – even if operating on the basis of an algorithm – then we can perhaps view the deployer as the one being fooled. An alternative position, and the one preferred by the current author, is that the provision is referring to a human, and that machines cannot be fooled into anything: they are just executing their programming. For situations where the TS has a human component, such as an operator who is brought to cancel a strike because an adversarial makes them believe the entity in question is protected, the argument could perhaps more easily be made that it does constitute perfidy, as an actual human is being fooled (instead of purely a model).

Additionally, the distinction made above between specific false negative and agnostic false negative may be useful here. With specific false negative, the adversary is knowingly forcing the TS to output a protected label, while with agnostic false negative, the adversary is not. The intent to abuse IHL would thus be easier to establish in the former situation compared to the latter.

We now briefly turn to the second matter of human shielding, which is "the use of persons protected by international humanitarian law … to deter attacks on

154 W. H. Boothby, above note 8, p. 385 fn. 57.

155 AP I, Art. 85(3)(f).

156 See Sean Watts, "Law-of-War Perfidy", *Military Law Review*, Vol. 219, 2014, p. 153. Simple perfidy is used by Watts to refer to actions that betray the opponent's confidence in IHL but which only lead to military or tactical benefit without causing any death, injury or capture. Cf. AP I, Arts 37(1) and 85(3)(f), which are result violations.

157 Yves Sandoz, Christophe Swinarski and Bruno Zimmerman, *Commentary on the Additional Protocols*, ICRC, Geneva, 1987, para. 1500.

combatants and military objectives".[158] One way in which an adversary can do this, as alluded to above, is by overwhelming the proportionality calculation so that the deployer cannot proceed without causing excessive incidental harm (or at least making the deployer believe this is the case).[159] For instance, the adversary could immunize its military compound by making a TS believe that the adversary's soldiers guarding the compound are civilians, or making the TS hallucinate dozens of civilians on the roof. Is this human shielding? The author's first instinct is to answer in the negative, since the adversary is not factually placing civilians in danger when executing this adversarial, which is ultimately the point of the prohibition.[160] To present an analogy: would the adversary be shielding if it had spread misinformation that dozens of civilians were residing in empty rooms adjacent to an arms depot, and the deployer were naive enough to believe this without checking further? However, arguments can reasonably be made for either position.

Concluding remarks

This article has moved away from the users of AI systems to focus on their opponents: adversaries. Adversaries are underappreciated actors in legal discussions on AI, but play a crucial role in actual combat situations. As the use of military AI expands, so too will the importance of adversarials. These powerful countermeasures potentially grant adversaries far-reaching influence over how an AI system behaves on a battlefield. If these systems are destructive, as assumed in this article, the protection of civilians from their effects becomes a primary concern. IHL encapsulates its most stringent protections into a set of obligations related to attack, but when the nature of adversarials makes it ambiguous or debatable which party is considered as effectuating an attack, the risk arises that neither party will uphold the associated obligations, ultimately making civilians pay the price.

To properly navigate this challenging subject matter, this article constructed an analytical foundation by exploring adversarials from both a technical and tactical perspective. This allowed us to see how adversarials differ from other countermeasures, and to understand the mindset of those utilizing them. Subsequently, analysis was performed on the legal criteria for an action to qualify as an attack, and the obligations associated with both attack and "sub-attack" situations. While adversarials differ from cyber attacks in many ways, it was nevertheless found that interpretative developments in the domain of cyber warfare, such as those in the Oslo Manual, are also useful and pertinent for legal analysis of adversarials. In particular, imperfect control, temporal and spatial distance between the adversary and the TS, and relative uncertainty as to how

158 Michael N. Schmitt, "Human Shields in International Humanitarian Law", *Columbia Journal of Transnational Law*, Vol. 47, No. 2, 2009, p. 293.
159 *Ibid.*, p. 298.
160 See AP I, Art. 51(7).

cyber attacks will cause first-order and second-order harms are also typical of many adversarials. These characteristics make it necessary to afford greater weight to concepts such as knowledge and foreseeability in order to determine when an adversary is exercising sufficient influence over the result so as to justifiably attribute said result to the adversary.

From this foundation, it was concluded that the adversary's mental state should be the dominant condition for determining an adversarial's status under IHL, and a framework was proposed primarily based on the different strategies outlined in the article. It was found that adversarials following some strategies (military false positive and civilian false positive) unquestioningly constitute attacks for which the adversary must take responsibility. For adversarials where direct intent is lacking (agnostic false positive and random drop), the assessment is casuistic: how foreseeable is it for the adversary that additional harm will result? In this regard, the article went further than the Oslo Manual: it was argued that the adversary should also incur attacker responsibilities in cases where it is *foreseeable* that its adversarial will render the TS indiscriminate. This balances the fact that many adversarials are objectively impossible for the deployer to detect before they manifest.

Finally, for the false negative strategy, the assessment depends on whether one would consider the adversarial as sufficient interference into the deployer's military activities in itself so as to constitute hostile action; and even if one's answer is yes, the nature of false negatives would not engender much conflict with obligations in attack for the adversary, since they render the TS less violent. Nevertheless, one may ask whether certain forms of false negatives may conflict with other prohibitions, such as those against human shielding and perfidy. Some remarks were offered on this subject, but these questions may be interesting to pursue further, particularly with more specific scenarios and adversarials.

The fact that one party is designated the attacker also does not absolve the other party of all obligations. If the adversary is considered the attacker, the deployer must still enact all measures to uphold the reliability and resilience of its systems and reduce risk to civilians. If the deployer is considered the attacker, the adversary is still bound by more general obligations, particularly the basic rule of distinction and constant care.

It is true that the framework's reliance on the adversary's mental state potentially produces difficulties for *post hoc* analysis, particularly for the purposes of determining State or individual responsibility after an incident. Particularly challenging is the "inherent inaccessibility of the internal facts".[161] It is impossible to objectively peer into a person's head and determine what they knew, wanted or had been thinking at a particular time. As Boothby concedes, such "fine distinctions are based on the intent of the [adversary], an intent which it is likely to be difficult to establish".[162] Having borrowed from criminal law

161 Keren Shapira-Ettinger, "The Conundrum of Mental States: Substantive Rules and Evidence Combined", *Cardozo Law Review*, Vol. 28, 2007, p. 2582.

162 W. H. Boothby, above note 6, p. 157.

theory to express the adversary's mental states in Table 1, however, we can also look to criminal law for a solution: because a person's thoughts are inaccessible, "the determination of a state of mind is to be deduced from evidence".[163] The evidence, in this case, would be the technical characteristics of the adversarial. For instance, a backdoor gives the adversary so much control over the TS's exact behaviour, and over the timing and context of when such behaviour is triggered, that the particular result was almost certainly willed by the adversary – i.e., it would be a military or civilian false positive, depending on the victim. This also highlights the importance of deployers keeping proper records and conducting post-deployment assessments. As the physical controller of the TS, the deployer is likely the only party placed to analyze what went wrong with its system, and accordingly, what type of adversarial the system was impacted by.

Finally, one may argue that in practice, evidence of the adversary's mental state will be hard to come by, because evidence is lacking, is classified or was willingly dissimulated.[164] This reality is hard to contest. However, recall that the principal concern offered in the introduction to this article was not *post hoc* investigation challenges, but lack of legal clarity: when operating in a fog, neither the deployer nor the adversary is certain as to who is required to implement obligations in attack. Thus, *good-faith adversaries and deployers* are the framework's main benefactors. Indeed, this article addresses itself to the good-faith adversary who pauses and asks: "Does putting on these adversarial sunglasses constitute an attack?" One hopes that this discussion brings illumination and a workable method for belligerents to determine whether IHL considers them as an attacking party. It is then up to the respective parties to take risk-mitigating measures appropriate to their role.

163 K. Shapira-Ettinger, above note 161, p. 2584.
164 For example, patterns can be quickly removed by the adversary after the adversarial succeeds.