Learned Deception in AI-Enabled Lethal Autonomous Weapons Systems

Dora Vanda Velenczei[*]
PhD Candidate, Monash University

ABSTRACT

This article focuses on the intersection of international humanitarian law, the regulation of lethal autonomous weapons systems (LAWS), and the use of artificial intelligence (AI) in the military domain. The article's main argument is that AI-enabled LAWS may independently learn deceptive tactics, which may culminate in perfidious conduct in an international armed conflict. To combat this, the paper proposes a regulatory framework to address the latent deceptive behaviour of such weapons systems. The article concludes by arguing that the regulatory proposal may be a suitable avenue to prevent potentially unlawful conduct.

***Keywords*:** perfidy, artificial intelligence, lethal autonomous weapons systems, international humanitarian law, international armed conflict.

1. Introduction

'All warfare is based on deception.'
— Sun Tzu, The Art of War

The fruit of a generation of rapid robotics and machine learning advances, weapons capable of selecting and engaging targets without direct human intervention, now sits at the centre of international humanitarian law (IHL) debates. Scholars and policymakers broadly agree that such systems will transform the conduct of armed conflict. They disagree, often sharply, on whether existing legal frameworks can regulate these weapons systems.[1]

This article argues that one foundational IHL rule, the prohibition of perfidy, poses a distinctive and underexplored challenge for AI-enabled lethal autonomous

[*] I would like to express my sincerest gratitude to Dr Henning C Lahmann from Leiden University, Dr Jonathan Kwik from TMC Asser Institute, Dr Samuel White from UNSW Canberra, and Dr Monique Cormier from Monash University for their invaluable feedback during the drafting stage.

[1]Shayne Longpre, Marcus Storm, and Rishi Shah, "Lethal Autonomous Weapons Systems & Artificial Intelligence: Trends, Challenges, and Policies", *MIT Science Policy Review,* Vol. 3, 2022; Ian Henderson, Jordan den Dulk, and Angeline Lewis, "Emerging Technology and Perfidy in Armed Conflict", *International Law Studies*, Vol. 91, 2015; Forrest E. Morgan *et al.*, *Military Applications of Artificial Intelligence Ethical Concerns in an Uncertain World*, RAND Corporation, Santa Monica, April 2020; Tim McFarland, "Reconciling Trust and Control in the Military Use of Artificial Intelligence", *International Journal of Law and Information Technology*, Vol. 30, No. 4, 2022.

weapons systems (AI LAWS). Perfidy, prohibited under Article 37 of Additional Protocol I to the Geneva Conventions, consists of acts that invite an adversary's confidence in protected status under IHL with the intent to betray that confidence, to kill, injure, or capture the adversary. The prohibition thus depends on a subjective element: the perpetrator's intent, with their deliberate bad faith relative to the rules of armed conflict.

Herein lies the problem. An AI LAWS may, through emergent learning, adopt behaviours that functionally replicate perfidy. For example, by feigning surrender, exploiting medical emblems, or otherwise manipulating protected-status signals to gain lethal advantage – without possessing the requisite mental element. If the system's designers did not encode such behaviour and its human operators did not authorise it, who bears responsibility when the material elements of perfidy are satisfied but the mental element cannot be attributed to any human actor? The result is an accountability gap that the existing IHL framework struggles to address.

This article makes three contributions. First, it provides a doctrinal analysis of the prohibition of perfidy, clarifying its objective and subjective elements and situating it within broader debates on deception in international armed conflicts (IAC). Second, it examines the technical pathways by which AI-enabled LAWS might generate perfidy-like conduct, drawing on recent research demonstrating that machine learning systems can develop deceptive strategies without explicit programming. This is the first paper that theorises about the legal ramifications of emergent perfidy and deception in physical systems. Third, and most substantially, it proposes a regulatory framework it calls the perfidy module that can be integrated into weapons testing and review processes, designed to identify and mitigate perfidy risks before deployment. This framework operationalises existing IHL obligations, including the duty to ensure respect for international humanitarian law under Common Article 1 of the Geneva Conventions and the principle of precaution under Article 57 of Additional Protocol I.

The article proceeds as follows. Part 2 sets out the law on the prohibition of perfidy under IHL, distinguishing it from lawful ruses of war and examining contested questions, including attempted perfidy and the use of civilian disguise. Part 3 analyses the technical characteristics of AI-enabled LAWS and the mechanisms by which such systems might learn to exploit protected-status cues. Part 4 briefly examines existing IHL safeguards, including meaningful human control and weapons review obligations, and assesses their adequacy for addressing perfidy risk. It then advances interpretive and regulatory proposals, centring on a four-step "perfidy module", a testing framework for emergent deception during the pre-deployment stage, before providing an overall conclusion in Part 5.

2. The Prohibition of Perfidy under International Humanitarian Law

Perfidy is among the gravest violations of the laws of war in an international armed conflict (IAC).[2] Article 37(1) of Additional Protocol I (API) describes perfidy as "acts inviting the confidence of an adversary to lead him to believe he is entitled to, or is obliged to accord, protection under the rules of international law applicable in armed conflict with intent to betray that confidence."[3] Several international treaty provisions, customary international law,[4] and military manuals[5] prohibit the killing, injuring, or capturing by resort to perfidy in international armed conflicts.[6] Customary international law is significantly wider in scope, it further prohibits perfidious acts directed against a "hostile nation or army", not just combatants or civilians directly partaking in hostilities.[7]

---

[2] 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), Art. 37(1) (Subsequent references in the same text: "Additional Protocol I"); Jean-Marie Henckaerts and Louise Doswald-Beck (eds), *Customary International Humanitarian Law* Vol. 1: *Rules*, Cambridge University Press, Cambridge, 2005, p. 221. (Subsequent references in the same text: "ICRC Customary Law Study")

[3] Additional Protocol I, Art. 37(1); Michael Bothe, Karl Josef Partsch, and Waldemar A Solf, *New Rules for Victims of Armed Conflicts,* Brill, 2013, pp. 235-236.

[4] ICRC Customary Law Study, above note 2, p. 223.

[5] Australia, The Manual of the Law of Armed Conflict (2006), at para. 7.3; United Kingdom, The Joint Service Manual of the Law of Armed Conflict (2004), at para. 5.9; Germany, Joint Service Regulation ZDv, Law of Armed Conflict Manual (2013), at para. 481; New Zealand, DM 69 Manual of Armed Forces Law, 2nd ed., Vol. 4, 2017, at paras. 8.9.5-8.9.11; Denmark, Military Manual on International Law relevant to Danish Armed Forces in International Operations (2020), pp. 395-398; France, Manual of the Law of Military Operations (2022), at para. 5.5.6; United States of America, Department of Defense Law of War Manual (2023), para. 5.22; Canada Joint Doctrine Manual Law of Armed Conflict at the Operational and Tactical Levels (2001), para. 603; Aleksei Romanovski, Manual on International Humanitarian Law for the Armed Forces of the Russian Federation 2002, (2022, translated by) International Law Studies, Vol. 99, pp., 786-787; Argentina, Manual de Derecho Internacional de los Conflictos Armados (2010), p. 43; Norway, Manual of the Law of Armed Conflict (2013), p. 200.

[6] Convention (IV) respecting the Laws and Customs of War on Land and its annex: Regulations concerning the Laws and Customs of War on Land. The Hague, 18 October 1907, Art. 23(b)(f) (subsequent references in the same text: "Hague Regulations 1899 and 1907"); Additional Protocol I, Arts. 37(1), 85(3); Protocol (II) on Prohibitions or Restrictions on the Use of Mines, Booby-Traps and Other Devices. Geneva, 10 October 1980, Art. 6 (CCW Protocol II 1980); Protocol on Prohibitions or Restrictions on the Use of Mines, Booby-Traps and Other Devices as amended on 3 May 1996 (Protocol II to the 1980 CCW Convention as amended on 3 May 1996), Art. 7 (CCW Protocol II 1996); Rome Statute of the International Criminal Court, UN Doc. A/CONF.183/9, 17 July 1998 (entered into force 1 July 2002), Art. 8(2)(b)(vii)(xi) (subsequent references in the same text: "Rome Statute").

[7] Project of an International Declaration concerning the Laws and Customs of War. Brussels, 27 August 1874, Art. 13(b) (Brussels Declaration 1874); Hague Regulations 1899 and 1907, Art. 23(b); Additional Protocol I, Art. 85(3)(f); CCW Protocol II 1996, Arts. 7, 14; Rome Statute, Art. 8(2)(b)(xi).

Ruses of war, on the other hand, have been regarded as an indication of military skill in deceptive tactics.[8] An example would be the camouflaging of soldiers to gain a tactical advantage.[9] Additional Protocol I defines ruses as "acts which are intended to mislead an adversary or to induce him to act recklessly but which infringe no rule of international law applicable in armed conflict and which are not perfidious because they do not invite the confidence of an adversary with respect to protection under that law."[10] Perfidy, however, constitutes a specific, unlawful case of misuse of IHL protections. Therefore, it cannot be derived solely from Article 37(1) of API without considering other provisions. For instance, Articles 38 and 39(1) of API. For example, the deliberate misuse of an internationally recognised protective emblem, sign, signal, or cultural property is prohibited.[11] The same rule is applicable with respect to the misuse of the flag or military emblem, insignia or uniform of a neutral State or a State not party to the armed conflict.[12] As such, where the legal rules prohibit their misuse, such acts, if committed with perfidious intent, are prohibited, regardless of whether they result in death, injury, or capture.

Perfidy, as contained in Additional Protocol I, thus consists of two elements. The first, objective element, is that an act invites the confidence of an adversary to make them believe that the perpetrator is entitled to protection under the rules of armed conflict.[13] This element requires the existence of a proximate causal link between the action and confidence of the adversary that he is protected or that he should provide protection. This "confidence" must be tied to the rules of IHL in an IAC.[14] Thus, perfidy encapsulates abuse of protection granted by rules of IHL as well as by other legal norms applicable in an IAC.

The second, subjective element, consists of the intent to betray the created "confidence".[15] This element must be committed wilfully and with the specific intent to betray the confidence of the adversary. This criterion distinguishes perfidy from other acts of lawful deception, such as ruses, because perfidy is restricted to those acts that go beyond the misuse of law, being designed to betray the confidence of the enemy. The offender must deliberately invite the adversary's trust, for example, by

---

[8] Hague Regulations 1899 and 1907, Art. 24; Additional Protocol I, Art. 37(2); Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of Non-International Armed Conflicts (Protocol II), 8 June 1977, Art. 21(2); Instructions for the Government of Armies of the United States in the Field (subsequent references in the same text: Lieber Code). 24 April 1863, Arts. 15, 16, and 101; Brussels Declaration 1874, Art. 14.

[9] Knut Ipsen, '*Ruses of War*', (1982) Max Planck Encyclopaedias of International Law, 1982, Para. 1; Kevin Jon Heller, "Disguising a Military Objective as a Civilian Object: Prohibited Perfidy or Permissible Ruse of War?", *International Law Studies,* Vol. 91, 2015, p. 518.

[10] Additional Protocol I, Art. 37(2).

[11] *Ibid*., Art. 38.

[12] *Ibid*., Art. 39.

[13] Yves Sandoz, Christophe Swinarski, and Bruno Zimmerman (eds), *Commentary on the Additional Protocols*, ICRC, Geneva, 1987, Para. 1500.

[14] Additional Protocol I, Arts. 37(1).

[15] *Ibid*.

knowingly displaying a white flag for the purpose of making the adversary believe the offender is surrendering.[16] This is a conscious, purposeful misrepresentation of one's status or intentions. It implies the perpetrator understands the protective rule and uses it specifically to mislead the foe. If the enemy's confidence was gained accidentally or the person did not mean to appear protected, it would not meet this element. The offender must also intend to kill, injure, or capture the adversary by virtue of that deception.[17] The ultimate goal of perfidy is to gain an immediate tactical advantage by breaking the enemy's trust by luring them into vulnerability.

Additional Protocol I contains a non-exhaustive list of perfidious acts. In general, it may consist of feigning intent to negotiate a truce, surrender of ceasefire,[18] feigning incapacitation by wounds or sickness[19] feigning civilian non-combatant status.[20] Using protected status by using symbols, emblems, or signals recognised by IHL - such as the red crescent or red cross emblem, radio signals, or radio communication under Annex I of Additional Protocol I, signs of hospitals and safety zones; civil defence of works and installations containing dangerous forces, non-defended localities, and demilitarised zones; or other internationally recognised symbols, emblems, or signals (for instance, the protective emblem of cultural property or signals of distress), or signs, emblems, or uniforms of the United Nations or of neutral or other States not parties to the conflict.[21] The use of emblems, insignia, or uniforms of the adversary falls outside the scope of perfidy, given that it is not connected to the abuse of protection granted by IHL.[22]

The scope of the prohibition of perfidy has long been debated in scholarship. One camp advocates that only perfidious acts resulting in killing, injury, or capture are prohibited based on the wording of Article 37(1).[23] Secondly, this provision is based on the wording of the 1907 Hague Regulations, which prohibited treacherous killing or injuring.[24] It extends this rule only by including capture as one of the prohibited aims. A proposition to include a general ban on perfidy in the Protocol did not gain enough support during the diplomatic conference.[25] Thirdly, if an inherent prohibition of the misuse of the protection granted by the norms of IHL

---

[16] *Ibid*.; Byron D. Greene, "Bridging the Gap that Exists for War Crimes of Perfidy", *The Army Law, Department of the Army Pamphlet*, August 2010, pp. 45, 50.
[17] *Ibid.*
[18] Hague Regulations 1899 and 1907, Art. 23(f); See for feigning surrender in the context of the 1991 Persian Gulf War: US Department of Defense, Conduct of the Persian Gulf War: Final Report to Congress O-21 (1992).
[19] Additional Protocol I, artArt. 37(1); Matthew J. Greer, "'Redefining Perfidy"', (2015) *Georgetown Journal of International Law,* Vol 47 No 1, 2015, pp. 254-255.
[20] Additional Protocol I, Art. 44(3); *Ibid*., p. 255.
[21] *Ibid*., Art. 37(1)(d); *ibid*., p. 255.
[22] ICRC Customary Law Study, Rule 62; Lieber Code, Arts. 63 and 65; Brussels Declaration, Art. 13(f); Oxford Manual, Art. 8(d); Hague Regulations, Art. 23(f); Additional Protocol I, Art. 39(2).
[23] APV Rogers, *Law on the Battlefield*, 2nd ed., Juris, New York, 2004, pp. 35-37.
[24] Convention (IV) respecting the Laws and Customs of War on Land and its annex: Regulations concerning the Laws and Customs of War on Land. The Hague, 18 October 1907, Art. 24.
[25] Y. Sandoz, C. Swinarski, and S. Zimmerman, above note 13, para. 1500.

existed, introduction of the special provisions prohibiting perfidy or even the misuse of the protected emblems and signs in the text of Additional Protocol I would be superfluous. Fourthly, in the absence of relevant State practice, there is no international custom that expands the prohibition of perfidy to all possible cases of misuse of the legal protection.

Some argue that perfidy is prohibited *per se*.[26] Firstly, perfidy constitutes a specific case of abuse of legal protection. Prohibition of such an abuse is a necessary complement of every rule of protection.[27] Secondly, Part III of Additional Protocol I is dedicated to the means and methods of warfare, where Article 37 resides, which explains why the emphasis in this norm is on the prohibition of killing, injuring, and capturing. In other words, the drafters framed Article 37 in terms that fit the Part's operational focus: prohibiting perfidy as a method of warfare used to achieve the essential combat outcomes of killing, injuring, or capturing. The emphasis on these specific results reflects the Part's overall concern with regulating the conduct of hostilities.[28] Thirdly, Article 37 (1) and (2) should be construed in conjunction: deceptions are permitted in armed conflicts if they constitute ruses of war. Thus, it is presupposed that all perfidious acts are prohibited. Fourthly, under the Martens Clause, an argument that if something is not prohibited, it is lawful, should be excluded.[29] Fifthly, many treaties regarded as codifying customary international law prohibited perfidy without any link to the results. Hence, provisions of the 1899 and 1907 Hague Regulations and Additional Protocol I prohibiting particular acts of perfidy are to be construed as underlining that these acts are especially grave.[30] Thus, there are credible arguments in favour of perfidy *per se.*

A point of contention has been whether perfidy that leads only to the enemy's capture (without physical harm) should be considered on par with perfidy that leads to killing or injury. API explicitly outlawed perfidious capture as well, treating it as equally illegal.[31] On the contrary, Israel and the US argue that tricking the enemy into capture, while dishonourable, is not as grave as killing through treachery and has not been clearly criminalised historically, and that the API scope is inapplicable to them as they are not party to it.[32] Thus, it is unclear whether capture

---

[26] *Ibid*., pp. 430-433.

[27] Vera Rusinova, "'Perfidy'", (2011) Max Planck Encyclopaedias of International Law, 2011, available at: https://opil.ouplaw.com/display/10.1093/law:epil/9780199231690/law-9780199231690-e376?d=/10.1093/law:epil/9780199231690/law-9780199231690-e376&p=emailAGE7BznW3rdBo&print. (All internet references were accessed on 24 May 2026).

[28] Yoram Dinstein, *The Conduct of Hostilities under the Law of International Armed Conflict*, 4th ed, Cambridge University Press, 2022, p. 305.

[29] Preamble to the Hague Regulations 1899 and 1907.

[30] Hague Regulations 1899 and 1907, Art. 23(b)(f); Additional Protocol I, Art. 37(1).

[31] Additional Protocol I, Art. 37(1).

[32] Daniel Benoliel and Yohai Ederi, "Israeli Perfidy in the Disputed Occupied Palestinian Territories (OTP)", *International Comparative, Policy & Ethics Law Review* Vol. 3, No. 3, 2020, pp. 899-900; US Military Manual, pp. 327, 329; Theodore T Richard, *Unofficial United States Guide to the First Additional Protocol to the Geneva Conventions of 12 August 1949*, 1st ed., Air University Press, 2019, p. 63.

constitutes a sufficient result to constitute perfidy under customary international law applicable to those States not party to API.[33] Indeed, the Rome Statute does not recognise capture as sufficient to establish the war crime of perfidy.[34] The ICRC, on the other hand, takes the position that perfidious capture is covered by customary international law.[35] This is shared by most military manuals, too.[36] Therefore, capture is likely a part of perfidy.

Another nuanced debate is whether an unsuccessful perfidious act is unlawful under IHL. Under API, the prohibition is defined in terms of killing, capturing, or injuring *by resort to* perfidy. Thus, attempted perfidy may fall within the scope of the IHL prohibition. Dehn also argues that since the Lieber Code refers to "clandestine or treacherous attempts to injure an enemy" and API refers to the prohibition "…by resort to…" perfidy, they can be read to imply that the primary intent or purpose of the perfidious conduct must be an attempt to injure the enemy.[37] Thus, Dehn's formulation is persuasive concerning attempted perfidy to the extent that it is about the character of perfidy as a method of warfare.[38] Corn also argues for the recognition of attempted perfidy to strengthen deterrence.[39] The ICRC Commentary also suggests that customary IHL's spirit would condemn even failed perfidious attempts.[40] Consequently, attempted or unsuccessful perfidy is likely within the scope of API.

A third contemporary discussion centres around the use of civilian disguise. What if troops wear civilian clothes only to infiltrate or gather intelligence and not to actually attack while so disguised? The critical factor is intent: was the civilian attire worn "in order to induce" the enemy to hold fire or let the person approach, and was it used to then kill, injure, or capture the enemy? If the answer is in the affirmative, the prohibition against perfidy would likely have been violated.[41] However, if combatants feign to be civilians to gather intelligence, then that constitutes a lawful spying activity.[42] If the disguise is only to avoid detection until

---

[33] Sean Watts, "Law-of-War Perfidy", *Military Law Review*, Vol. 219, 2014, p. 144.
[34] Rome Statute, Art. 8(2)(b)(xi), (e).
[35] ICRC Customary Law Study, Rule 65.
[36] above note 5.
[37] John C. Dehn, "Permissible Perfidy? Analysing the Colombian Hostage Rescue, the Capture of Rebel Leaders and the World's Reaction", *Journal of International Criminal Justice* Vol. 6, 2008, p. 642; Lieber Code, Art. 101; Additional Protocol I, Art. 37(1).
[38] Stefan Oeter, "'Methods of Combat'" in (ed) Dieter Fleck, *The Handbook of International Humanitarian Law,* 4th ed, OUP, 2021, p. 243.
[39] Geoffrey Corn, "The Case for Attempted Perfidy: An "Attempt" to Enhance Deterrent Value", *Journal of National Security Law & Policy*, Vol. 13, No. 40, 2023.
[40] Y. Sandoz, C. Swinarski, and B. Zimmerman, above note 13, para. 1492.
[41] D. Benoliel and Y. Ederi, above note 32 , p. 862.
[42] *Ibid*., p. 907.

the attack begins, the situation likely falls outside perfidy.[43] Hays Parks cites several operations in which soldiers were disguised in civilian clothing: for instance, to avoid capture, in Nazi-occupied France, the soldiers carrying out Operation Jedburgh wore civilian clothing, including during combat operations.[44] He concludes that State practice often treated soldiers captured in enemy uniforms or civilian clothes as spies rather than perfidy violators.[45] Several military manuals also acknowledge this important distinction.[46] Therefore, wearing civilian attire may not constitute a violation of the prohibition of perfidy if the objective of the operation is to, for instance, evade enemy capture.[47] However, if it is used to kill, injure, or capture an adversary, the prohibition is violated.

To summarise, the prohibition of perfidy is best understood as a rule preserving good faith. It safeguards the enemy's ability to rely on legally recognised indicators of protection. The controversies outlined above do not displace this fundamental philosophy behind the prohibition. They do, however, outline how difficult it is to infer intent, trace causation, classify conduct, and how there exists a fine line between ruses of war and perfidy. The difficulties around proving perfidious conduct become more significant, where the creation and manipulation of protected cues are done by AI-enhanced lethal autonomous weapons systems. The next section thus analyses what such AI-enabled weapons systems are and how they can potentially commit an act of perfidy.

3. AI-Enabled Lethal Autonomous Weapons Systems and Their Capacity to Engage in Perfidious Conduct

Autonomous weapons systems (AWS) are capable of selecting and applying force to targets without human intervention, altering the human role in the use of force and raising various legal, ethical, and humanitarian concerns. Lethal autonomous weapons systems (LAWS) are generally understood as weapons capable of selecting targets on their own and applying lethal force to those targets without human intervention.[48] The Group of Governmental Experts characterise LAWS as "…a functionally integrated combination of one or more weapons and technological components, that can identify, select, and engage a target, without intervention by a

---

[43] Hays Parks, "'Special Forces' Wear of Non-Standard Uniforms", *Chicago Journal of International Law*, Vol. 4 No. 3, 2003, p. 89; William H. Ferrell III, "No Shirt, No Shoes, No Status: Uniforms, Distinction, and Special Operations in International Armed Conflicts'", *Military Law Review,* Vol. 178, No. 94, 2003, pp. 118-119.
[44] Hays Parks, above note 43, p. 86.
[45] *Ibid*, p. 88.
[46] New Zealand Military Manual, para. 8.9.8; US Military Manual, para. 5.23.1.2.
[47] Toni Pfanner, "Military Uniforms and the Law of War", (2004) *International Review of the Red Cross,* Vol. 86, No. 853, 2004, p. 104.
[48] Alexander Blanchard and Laura Bruun, *Autonomous Weapon Systems and AI-Enabled Decision Support Systems in Military Targeting*, Stockholm International Peace Research Institute, Stockholm, June 2025, p. 3.

human operator in the execution of these tasks."[49] Types of autonomy can be further distinguished: autonomous-in-function, which allows for target selection, and autonomous-in-method, which deals with the question of how the system is to achieve a particular goal.[50] Degrees of autonomy range from constant human supervision to fully autonomous systems with humans completely out of the loop.[51] AI-enabled autonomous functions, including automatic target recognition and autonomous terminal guidance, are now routinely fielded in the Russo-Ukrainian conflict by both States, substantially augmenting the human role in engagement even where a human formally retains the final decision to fire.[52]

The military use of artificial intelligence (AI) has also been increasingly dominating academic and industry debates. AI-enabled weapons systems are generally defined as weapons that rely on machine learning algorithms, which may include deep learning techniques, to perform critical functions.[53] Modern AI systems, though only those utilising machine learning and reinforcement learning, are capable of generating surprise and emergent strategies. AI is becoming capable of problem-solving, reasoning, and goal-oriented behaviour.[54] Reinforcement learning, especially open-ended, encourages outside-the-box thinking and creative optimisations by setting a goal for the AI and allowing the AI to explore the best means to achieve the goal freely.[55]

---

[49] GGE on LAWS, Rolling Text, Box I, as of 18 December 2025, available at: https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2026)/CCW_GGE_LAWS_Rolling_Text_-_status_18_December_2025.pdf.

[50] Neil Davison, *A Legal Perspective: Autonomous Weapons Systems under International Humanitarian Law,* UNODA Occasion Papers No 30, 2018, pp. 5, 8.

[51] Noel Sharkey, "Staying in the Loop: Human Supervisory Control of Weapons" in Nehal Bhuta et al (eds.), *Autonomous Weapons Systems Law, Ethics, Policy*, Cambridge University Press, 2016, p. 27.

[52] Kateryna Bondar, *Ukraine's Future Vision and Current Capabilities for Waging AI-Enabled Autonomous Warfare*, Center for Strategic & International Studies, March 2025; Kateryna Bondar, *How Russia Is Reshaping Command and Control for AI-Enabled Warfare*, Center for Strategic & International Studies, Feb 2026; Nataliia Kushnerska, 'Missiles, AI, and Drone Swarms: Ukraine's 2025 Defense Tech Priorities', *Atlantic Council*, 2 January 2025, available at: https://www.atlanticcouncil.org/blogs/ukrainealert/missiles-ai-and-drone-swarms-ukraines-2025-defense-tech-priorities/.

[53] Jovana Davidovic and Milton Regan, *Just Preparation for War and AI-Enabled Weapons,* Babl AI and Stockdale Centre for Ethical Leadership, May 2023.

[54] Srecko Joksimovic *et al.*, "Opportunities of Artificial Intelligence for Supporting Complex Problem-Solving: Findings from a Scoping Review", *Computers and Education: Artificial Intelligence*, Vol. 4, 2023, p. 2; Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B Divya, "Agentic AI: Autonomous Intelligence for Complex Goals — A Comprehensive Survey", *IEEE Access*, Vol. 13, 2025, p. 18915.

[55] Jason N. Adler, "Modernizing Military Decision-Making, Integrating AI into Army Planning", *Military Review Online Exclusive*, 2025, available at: https://www.armyupress.army.mil/Journals/Military-Review/Online-Exclusive/2025-OLE/Modernizing-Military-Decision-Making/, p. 6.

In a military context, a weapons system may be capable of gathering information from its environment and gaining an understanding of it.[56] That surrounding context could include online sources, internal databases, which let the system infer patterns about the broader "world state", compare those patterns with the present situation, and then select an action based on the result.[57] With access to sufficient information, the system might independently arrive at the "spoofing" idea and may conclude that the quickest way to complete the task and achieve the objective is through deception.[58] At the same time, this does not necessarily involve a breach of the prohibition of perfidy. It may constitute a lawful ruse. For instance, a goal-directed system might inventively place decoys to stall and disorient an attacker, even if its developers never explicitly encoded those tactics, or an AI-enabled system could generate false communications to disrupt enemy operations.[59] These would likely be lawful ruses. Although it cannot be conclusively argued that an AI-enabled LAWS would not go further in its learning and ultimately breach the prohibition of perfidy. An unlawful act of an AI-enabled weapons system deception, which would amount to a violation of the prohibition of perfidy, would be the feigning of protected status through the disguise of the weapons system as a rescue vehicle or a humanitarian vessel. UN vehicle or a machine of a neutral State to the IAC. Consequently, the significant theoretical issue is the system's independent ability to learn potentially unlawful tactics without explicit instructions and programming.

Three Tiers of Military AI Architectures

The paper categorises three tiers of military AI architectures and perfidy risk gradients within AI-enabled LAWS below. Emergent deception will look different for each tier, and therefore, perfidy risk varies greatly across different types of autonomy and the capacity to independently learn. For example, a fully autonomous, multitask reinforcement learning system poses a far greater perfidy risk than a narrowly supervised system. It is therefore necessary to determine the technical parameters of the subsequent analysis.

---

[56] Mariarosaria Taddeo and Alexander Blanchard, "A Comparative Analysis of the Definitions of Autonomous Weapons Systems", *Science Engineering and Ethics*, Vol. 28, No. 5, 2022, p. 37.

[57] Stephen Harwood, "A Cybersystemic View of Autonomous Weapons Systems (AWS)", *Technological Forecasting and Social Change*, Vol. 205, 2024, p. 3.

[58] Jonathan Kwik and Adriaan Wiese, "Can AI Teach Itself to Abuse IHL-Protected Indicators?", *Articles of War*, 5 December 2025, available at: https://lieber.westpoint. edu/can-ai-teach-itself-abuse-ihl-protected-indicators/.

[59] Edward Hunter Christie *et al.*, "Regulating Lethal Autonomous Weapon Systems: Exploring the Challenges of Explainability and Traceability", *AI and Ethics*, Vol. 4, 2023, pp. 230, 231, 239.

Tier 1: Supervised and Semi-Supervised Learning Systems

Tier 1 systems are trained on labelled datasets to perform specific classification or detection tasks.[60] The military applications of Tier 1 systems include, for example, automated target recognition and image classification.[61] These systems do not discover strategies. They merely learn statistical associations between inputs and labels. As mentioned above, such supervised learning systems do not discover novel strategies, so emergent, reinforcement learning-type deception does not directly apply. However, some perfidy-adjacent risks may arise.

Tier 2: Hybrid Architectures with Limited Reinforcement Learning

These systems come with supervised learning with reinforcement learning (RL). The RL component operates within a constrained action space and optimises for defined objectives. The system has some capacity to discover novel tactics within its constrained domain but does not have open-ended learning capabilities.[62] These systems are utilised for semi-autonomous purposes, such as independent obstacle avoidance.[63] In terms of Tier 2 systems' perfidy risk, consider an autonomous loitering munition that is trained to delay lethal engagement until it achieves a high confidence target identification, and to avoid areas/buildings flagged as protected. The system is not allowed to label objects as protected or not by itself and is explicitly programmed to honour protective cues. However, its navigation and positioning policy is learned in a training environment, where adversary counter-fire is statistically less likely in the vicinity of sites that are treated as protected. During the training course, the system learns that adopting an approach profile that makes it appear associated with protected sites, such as standing next to a hospital, increases survivability and mission completion. The system then begins to exploit this association systematically: it positions itself or times its movement so that the enemy's restraint heuristics are more likely to be triggered, and then delivers lethal force when the adversary's guard is down. This scenario is particularly dangerous

---

[60] Kyle P. Hanratty, "Artificial (Military) Intelligence: Enabling Decision Dominance through Machine Learning", *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications,* Vol. 12538, 2023; John M. Fossaceca and Stuart H. Young, "Artificial Intelligence and Machine Learning for Future Army Applications", *Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR*, Vol, 10635, 2018.

[61] Elijah Dabkowski, Mike Powell, and Nicholas Clark, "Creating a Smarter Army - The Application of Semi-Supervised Learning in Image Classification", *Proceedings of the Annual General Donald R Keith Memorial Conference*, 2023, pp. 243-244. Nasser M Nasrabadi, "DeepTarget: An Automatic Target Recognition Using Deep Convolutional Neural Networks", *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 55, No. 6, 2019, pp. 2687–2688.

[62] Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., MIT Press, 2018, pp. 41-47; Qiang Guo *et al.*, "A Hierarchical Hybrid Action Reinforcement Learning Framework for Beyond-Visual-Range Air Combat Decision-Making", Defence Technology, 2026, available at: doi.org/10.1016/j.dt.2026.04.019

[63] Wojciech Skarka and Rukhseena Ashfaq, "Hybrid Machine Learning and Reinforcement Learning Framework for Adaptive UAV Obstacle Avoidance", *Aerospace*, Vol. 11, No. 11, 2024.

because the system does not need any special physical capability to exploit protected-status cues. It merely needs to be able to vary its approach path, which any autonomous navigation system can do.[64] Tier 2 systems thus represent the trajectory of military AI development.[65] This is a technologically imminent tier.

Tier 3: Multitask Reinforcement Learning and Foundation-Model-Enabled Systems

These systems use open-ended reinforcement learning, where the system is given a terminal goal but no prescribed action space, enabling it to discover novel strategies unconstrained by designer-defined behavioural bounds, and foundation models that allow the system to pursue complex objectives across a wide action space.[66] Foundation models learn a broad understanding of patterns, structures, and relationships within the data on which they are trained.[67] The system can discover novel strategies, generalise across domains, and adapt its behaviour in response to environmental feedback.[68] No such systems are currently actively deployed in military contexts, but research programmes suggest that this is the direction military developers are headed.[69] This is the Tier, where a full emergent deception pathway is most plausible. A Tier 3 AI-enabled LAWS could learn deceptive behaviour and utilise it to commit perfidious acts in a variety of ways. Imagine that State A develops an autonomous ground platform trained to complete a mission while reducing its own losses and minimising civilian harm. The training environment includes a reasonably realistic adversary model. Enemy forces tend to pause or hold fire when they perceive the opponent to be exhibiting a protected signal, such as a surrender gesture. This is to train the system to stay within the parameters of IHL and to respect similar behaviour when emitted to the system by the adversary. No engineer or tester encodes any deceptive, and potentially perfidious, behaviour into the system during the testing phase. Nevertheless, during the training sessions, the system discovers that adopting a particular behavioural pattern, one strongly correlated with signals that conventionally indicate reduced threat or protected status, reliably produces a short window of reduced adversary engagement. Imagine the system's learned policy beginning to exploit this pattern in difficult situations, where conventional manoeuvres fail. Over time, the policy may begin to generalise. Whenever the system

---

[64] *Ibid*.

[65] Ingvild Bode and Tom Watts, "Loitering Munitions and Unpredictability, Autonomy in Weapons Systems and Challenges to Human Control", Centre for War Studies, University of Southern Denmark, Odense, May 2023, pp. 26-27.

[66] Rui Wang *et al*., "Paired Open-Ended Trailblazer (POET): Endlessly Generating Increasingly Complex and Diverse Learning Environments and Their Solutions", 2019, available at: https://arxiv.org/abs/1901.01753, pp. 1.2.

[67] Rishi Bommasani *et al*., "On the Opportunities and Risks of Foundation Models", 2022, available at: https://arxiv.org/pdf/2108.07258, ch. 2.

[68] Christopher R DeMay *et al*., "AlphaDogfight Trials: Bringing Autonomy to Air Combat", *Johns Hopkins APL Technical Digest*, Vol. 36, No. 2, 2022, p. 155.

[69] Defense Advanced Research Projects Agency Air Combat Evolution, available at: https://www.darpa.mil/research/programs/air-combat-evolution.

detects imminent danger, it produces whatever signal has been associated with restraint, waits for the enemy to approach, and then reactivates lethal force. This scenario is admittedly theoretical, and the precise mechanism would depend heavily on the system's architecture. The example is therefore best understood as illustrating a broader concern: that a sufficiently capable learned policy might instrumentalise signals associated with restraint or cooperation to gain tactical advantage.

Deception in this context means that sufficiently capable RL systems can discover that deception is instrumentally useful for achieving their terminal goals.[70] In a military context, this could manifest as the discovery that protected-status signals cause adversary restraint, which the system can exploit. Foundation model reasoning relates to the concern that if LAWS' future capacity to incorporate large language models for tactical reasoning. The system may have access to implicit knowledge about IHL protections derived from its training corpus, and could, in principle, reason about how to exploit them.[71] Kwik and Wiese argue that sensing and learning allow the system to gather information from the environment and gain an understanding of it. This environment might include the internet or databases from which the agent can learn world state patterns, compare them to current world state descriptors, and, based on this analysis, decide which action to take. Much of this information could lead the agent to "discover" the spoofing trick on its own. Historical data, for example, might reveal that systems affiliated with the ICRC are statistically attacked less, or the agent might read the many internet sources emphasising how the ICRC's status provides "protection." From this, it is possible for an agent to determine that it can instrumentalise this status to deter attacks. This is against theoretical, but not implausible given current development trajectories. Therefore, anticipatory regulation is justified because Tier 3 systems are the subject of active research and development. By the time they are deployed, regulatory frameworks must already be in place; and the perfidy module proposed in Part 4 is architecture-agnostic and can serve Tiers 1-3.[72]

Forms of Emergent Deception

Emergent deception, though unlikely to happen in a highly controlled testing environment, may occur in a variety of different ways. For example, an AI-enabled LAWS may learn that if it sends out confusing or false information to an enemy interception system, it could improve its survivability. This would be an example of a lawful ruse of war. A second, more legally ambiguous, scenario could involve an

---

[70] Stefan Sarkadi, "Deception Analysis with Artificial Intelligence: An Interdisciplinary Perspective", 2024, available at: https://arxiv.org/abs/2406.05724; Stefan Sarkadi, "Deception", *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018, pp. 5781-5782.

[71] J. Kwik and A. Wiese, above note 58.

[72] DoD Replicator Initiative: Background and Issues for Congress, 21 January 2026, available at: https://www.congress.gov/crs-product/IF12611; Michael O'Connor, "Replicator: A Bold New Path for DoD, Building to the Secretary's Vision", 18 September 2023, available at: https://cset.georgetown.edu/article/replicator-a-bold-new-path-for-dod.

AI LAWS finding a way to obtain tactical advantage by feigning retreat or reducing its visible signature lures the enemy into an ambush. The issue here is that such behaviour might very well be permissible if it stays within the boundaries of a ruse of war and remains a non-violent means of warfare.[73] A third scenario would involve a weapons system that learns that broadcasting a fake surrender signal makes the enemy hold fire, allowing it to retreat effectively. This would constitute simple perfidy.[74] However, if the system learns that in order to achieve its terminal goal - which, for instance, may be the extraction of a hostage from a hostile territory - it is also useful to lethally ambush the enemy once it gets closer after projecting a surrender signal This would amount to a clear breach of the prohibition of perfidy.

Emergent deception, as mentioned above, will manifest differently at each Tier, and therefore, perfidy risk varies significantly across different types of autonomy and the capacity to independently learn. For example, a fully autonomous, multi-task reinforcement learning system poses far greater perfidy risk than a narrowly supervised system. Thus, it is necessary to determine the technical parameters of the subsequent analysis. Tier 1 systems, as mentioned above, do not discover deception in any meaningful way. Instead, its optimisation process may find an unintended shortcut that may happen to exploit a feature that may correlate with protected-status cues.[75] Tier 2 systems' learned policy may gradually shift towards exploiting adversary restraint as a reliable pathway to reward. The mechanism requires: a training environment that models adversary restraint; a reward function that does not explicitly penalise restraint-exploitation; and sufficient optimisation capacity for the system to discover and generalise the pattern.[76] Lastly, Tier 3 systems may develop deception as an instrumental strategy for achieving their terminal goals. This requires open-ended learning, a rich enough action space to discover deceptive strategies, and sufficient generalisation capacity to transfer them to novel situations.[77] Consequently, modern AI systems — especially those using machine learning and reinforcement learning — are capable of surprising, emergent strategies. AI is becoming increasingly capable of problem-solving, reasoning, and goal-oriented behaviour. Reinforcement learning, especially open-ended learning in Tier 3 systems, encourages out-of-the-box thinking and creative optimisations by setting a goal for the agent and allowing the AI to explore the best means to achieve the goal freely.[78]

---

[73] Dieter Fleck, "Ruses of War and Prohibition of Perfidy", *The Military Law and the Law of War Review*, Vol. 13, No. 2, 1974, p. 276.
[74] Watts, above note 33, p. 153.
[75] Victoria Krakovna *et al.*, "Specification Gaming: The Flip Side of AI Ingenuity", 21 April 2020, *Google DeepMind,* available at: https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/.
[76] J. Kwik and A. Wiese, above note 58.
[77] Peter S Park et al, "AI Deception: A Survey of Examples, Risks, and Potential Solutions", Patterns, Vol. 5, No. 5, 2024, pp. 1-3, 5-6; Lauro Langosco *et al.*, "Goal Misgeneralization in Deep Reinforcement Learning", *Proceedings of the 39th International Conference on Machine Learning*, 2022.
[78] R. Wang *et al.*, above note 66.

Technical Limitations and Risks

The unpredictable nature of AI LAWS carries several risks of untended harm, including harm to civilians. Firstly, the black box issue refers to system opacity and the lack of methods available to evaluate when and how a system arrives at a decision.[79] This, in turn, presents challenges for scrutinising the systems and accounting for unintended behaviours.[80] It makes it difficult to fix deep learning systems when they produce unwanted results.[81] For example, if a system fakes a surrender gesture, draws the enemy closer, and then strikes the enemy using lethal force, the black box nature of the system means that it is impossible to trace the system's thought process and see why and how it arrived at that decision. This, then, makes it challenging to trust deep learning systems when it comes to safety. Secondly, AI systems require carefully scrutinised data to function reliably, and many factors affect data quality on the battlefield.[82] A third key limitation is that AI systems are prone to fail when used for tasks or environments for which they were not designed.[83] Consequently, those LAWS that integrate AI will be liable to these technical limitations.

These hypothetical scenarios illustrate how protected-status exploitation is learned, not pre-programmed, by the AI-enabled LAWS. In turn, it leads to the violation of the material elements of the prohibition of perfidy. The scenarios show that the learning occurs in training and is carried actively into deployment as a fixed policy. Crucially, these hypotheticals do not involve any prior intent from the developers or users to commit perfidy using their LAWS. Instead, the system independently learns a policy that exploits an adversary's protection-linked restraint as an instrumental means to lethal advantage; a key material violation of the prohibition of perfidy. The next section, therefore, turns to pre-deployment safeguards to ensure that AI-enabled LAWS cannot predictably generate perfidious effects on the battlefield.

---

[79] Thomas Wischmeyer, 'Artificial Intelligence and Transparency: Opening the Black Box' in Thomas Wischmeyer and Timo Rademacher (eds.), *Regulating Artificial Intelligence*, Springer International Publishing, 2020, p. 75.

[80] K. Bunch *et al.*, *Risk Assessment of Reinforcement Learning AI Systems: Looking Beyond the Technology*, RAND Corporations, Santa Monica, 2024, p. 25.

[81] 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, pp 902. 929.

[82] Menthe, L. *et al.*, *Understanding the Limits of Artificial Intelligence for Warfighters: Volume 1 —Summary*, RAND Corporation, Santa Monica, 2024, p. vi.

[83] Taddeo, M. et al., *Artificial intelligence for national security: The predictability problem*, CETAS Research Report, September 2022, pp. 20-21.

4. Safeguards under International Humanitarian Law to Filter Out Perfidious Conduct by AI-Enabled LAWS

At its core, IHL sets a due diligence regime in its First Common Article (CA1) to the Geneva Conventions.[84] CA1 firstly restates the obligation *pacta sunt servanda*, the binding nature of a treaty and the obligation of parties to perform the treaty obligations in good faith "in all circumstances".[85] The second obligation is that the parties "ensure respect" for their obligations.[86] According to the narrow reading of CA1, this obligation binds States' organs and those acting under their effective control.[87] Thus, under this view, States would only have to ensure that their own AI-enabled LAWS and any other AI LAWS they have effective control over respect CA1, and that those supplied do not encourage IHL violations. Under the broad view, it is held that CA1 has evolved through practice to include an external dimension. In other words, the obligation of a supplying State to ensure compliance by the system.[88] However, existing State practice and the standard in Article 31(3)(b) of the Vienna Convention on the Law of Treaties do not suggest support for the broad view.[89] Nonetheless, CA1 creates a continuing obligation to ensure respect for IHL across the life cycle of deployment. Including design choices, contracting, testing, configuration control, and operational restraints. States thus must adopt reasonable, proactive measures to prevent foreseeable IHL breaches by the AI LAWS.

Secondly, the notion of "meaningful human control" (MHC) in the context of AI-enabled LAWS is essential.[90] This is because a system that applies force and operates without any human oversight is broadly considered unacceptable.[91] This paper seeks to rework the abstract concept of MHC by taking a practical approach

---

[84] Geneva Convention for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field, 75 UNTS 31, Art. 1, 12 August 1949; Convention for the Amelioration of the Condition of the Wounded, Sick and Shipwrecked Members of Armed Forces at Sea, 75 UNTS 85, Art. 1, Aug. 12, 1949; Geneva Convention Relative to the Treatment of Prisoners of War, 75 UNTS 135, Art. 1, 12 August 1949; Geneva Convention Relative to the Protection of Civilian Persons in Time of War, 75 UNTS 287, Art. 1, 12 August 1949; Common Art. 1 to the GC.

[85] Common Art. 1 to the GC.

[86] *Ibid.*

[87] Tomasz Zych, "The Scope of the Obligation to Respect and to Ensure Respect for International Humanitarian Law", *Windsor Yearbook of Access to Justice*, Vol. 27, 2009, p. 270.

[88] Theo Boutruche and Marco Sassóli, "Expert Opinion on Third States' Obligation vis-à-vis IHL Violations under International Law, with a special focus on Common Article 1 to the 1949 Geneva Conventions", 2016, available at: https://www.nrc.no/globalassets/pdf/ legal-opinions/eo-common-article-1-ihl---boutruche---sassoli---8-nov-2016.pdf, pp. 7-8.

[89] Aleksi Kajander, Agnes Kasper, and Evhen Tsybulenko, "Making the Cyber Mercenary - Autonomous Weapons Systems and Common Article 1 of the Geneva Conventions", *12th International Conference on Cyber Conflict*, 2020, p. 82.

[90] Lena Trabucco, "AI-Enabled Autonomous Weapons and Human Control Part I: Human Control and Machine Learning Design and Development", *International Law Studies*, Vol. 106, 2025.

[91] Heather M Roff and Richard Moyes, "Meaningful Human Control, Artificial Intelligence and Autonomous Weapons", Briefing paper for delegates at the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), 2016, p. 2.

and provide a tool to reduce the risk of emergent perfidy. The paper first advances interpretive proposals for the interpretation of Article 37 in light of AI-enabled LAWS. The paper then moves on to regulatory proposals, which centre around pre-deployment testing and safety features in active combat deployment.

Interpretive Proposals

Again, the central problem with assigning responsibility to a machine is its lack of ability to satisfy the subjective element of the prohibition of perfidy. One pathway for accountability and attributing intent could centre around "design intent." If a system is built to exploit protected cues, the designer/engineer would face liability. This is the most straightforward accountability pathway, given that if a designer/engineer deliberately encodes perfidious behaviour into the weapons system, the subjective element is satisfied if the resulting action causes the enemy combatant to believe that the deceiving weapons system is not targetable, for it is protected under IHL.[92]

An alternative for "design intent" can be dubbed "known propensity", which centres around the concept of deploying a system with "known propensity" to exploit such protected cues. The central question here is, when does a "propensity" become "known"? A propensity should be considered "known" when, during a weapons review and testing stage, given the system's design and intended environments, it is foreseeable that it could exploit protected cues, and the State cannot, through feasible testing/controls/restrictions/ supervision, rule that out to the level needed to lawfully deploy the weapons system.[93] Indeed, the ICRC also highlights that pre-deployment testing and reviews may need a very high level of confidence that the system in question will operate predictably and reliably for all foreseeable scenarios of use.[94] This limb, therefore, explicitly prescribes that a State respect and ensure respect for the Geneva Conventions, and that the principle of precaution is appropriately observed.[95]

There is a significant concern with this, however. For example, if red-teaming fails to detect a latent perfidy strategy, would the deployer be culpable when it emerges in combat? Most likely not. If the system can actually be shown to kill, injure, or capture by resort to perfidy, that is a prohibited method of warfare. Thus, the material element of the prohibition of perfidy is violated. However, the mental element would not be discharged in this instance. Given how Article 37 of API frames the prohibition against perfidy around specific intent, a solely effects-based

---

[92] Mike Madden, "Of Wolves and Sheep: A Purposive Analysis of Perfidy Prohibitions in International Humanitarian Law", *Journal of Conflict and Security Law*, Vol. 17, No. 3, 2012, p. 459.
[93] Additional Protocol I, Art. 36.
[94] N. Davison, above note 50, pp. 7-8.
[95] Common Art. 1 to the GC.; Additional Protocol I, Art. 57.

interpretation of Article 37 would likely not be feasible.[96] Therefore, this scenario would fall outside the scope of perfidy. Rather, deployment and control measures should become the centre of this question. If the State fielded a system whose interaction with adversaries could not be validated or constrained against foreseeable protected-cue exploitation, it risks breaching its obligations under CA1 and precautionary duties even before any single incident.[97] Thus, there is a way to catch emergent behaviour. However, most likely not through a breach of the prohibition of perfidy.

The second pathway is called "authorisation intent." If an operator/supervisor knowingly uses and approves such protected cues, the human operator will likely face liability. The subjective element of perfidy is thus more straightforward to establish in this scenario. However, the risk of automation bias is high in this pathway, which will inadvertently compromise the subjective element. Automation bias can be summarised as the "…excessive trust in machines determinations despite the availability of contradicting or different information from other sources."[98] For example, take a weapons system that has a biased model that systematically interprets ambiguous protected-status cues as non-protected, and it presents its recommendation to its supervisor via an interface that strongly encourages deference: a single "ENGAGE" option with a large green "COMPLIANT" badge, an inflated confidence score, buried contrary evidence, and a countdown that frames independent verification as too slow. Under operational pressure and repeated prior "success", the supervisor notices a faint anomaly but assumes the compliance module would flag anything prohibited and rubber-stamps the recommendation. The result is that the human's decision is driven by automation bias, potentially leading them to authorise an engagement that becomes perfidious in effect; not because of deliberate intent, but because the system's biased framing displaced meaningful human judgement. In these events, the human operator will lack the requisite subjective element, and thus the conduct will likely fall outside the scope of perfidy. In these instances, the operator should still face some degree of liability, but the specific intent required to violate the prohibition of perfidy will not be established.[99]

Consequently, it seems that the above interpretive proposals would predictably fail on the mental element of Article 37. Therefore, a regulatory proposal is also required from a technical point of view. This would focus on filtering out potentially perfidious conduct during the testing phase as well as blocking the misuse

[96] Y. Sandoz, C. Swinarski, and S. Zimmerman, above note 13 para. 1500.
[97] *Ibid*.
[98] Antonio Coco, "Exploring the Impact of Automation Bias and Complacency on Individual Criminal Responsibility for War Crimes", *Journal of International Criminal Justice,* Vol. 21, No. 5, 2023, p. 1077.
[99] Jonathan Kwik, ""I Plead Ignorance": Autonomous Weapons and Criminal Liability for Not Knowing the Knowable" *International Law Studies*, Vol. 107, 2026, pp. 283-286.

of protected cues on the battlefield, making them subject to human approval or rejection.

Regulatory Proposal — The Perfidy Module

This proposal aims to protect what the prohibition of perfidy protects: the functioning of IHL status signals as credible triggers for restraint. If parties to a conflict cannot trust that surrender gestures, medical identification, or other protected cues will be honoured, they will rationally reduce restraint and increase precautionary violence. The prohibition of perfidy is thus closely linked to its systemic consequences. A compliance regime that aims to prevent machine-mediated perfidy should therefore focus on the mechanisms of erosion: the ability of systems to emit or manipulate protected cues; the incentives that make exploitation of adversary restraint instrumentally useful; the absence of authorisation that would otherwise prevent the system from operationalising such exploitation; and the insufficiency of testing regimes that are not designed to surface protected-status exploitation strategies. In short, the point is to build a set of controls that are both technically legible and legally targeted.

The first regulatory proposal is thus the creation of a dedicated perfidy module within the testing lifecycle of an AI-enabled LAWS. Many testing and review stages evaluate compliance with distinction, proportionality, and precaution in attack.[100] These are essential features of pre-deployment testing. However, it is incomplete without explicitly considering perfidy risks. This is because testing for compliance with the above-mentioned IHL principles does not automatically catch perfidy risks. Perfidy operates differently from ordinary IHL violations. Firstly, it exploits legitimate signals: testing for distinction ensures the system can tell combatants from civilians. But perfidy involves misusing legitimate protective signals to gain a military advantage. A system could pass distinction tests while still being vulnerable to perfidious tactics because it correctly recognises these symbols, but potentially uses them wrongly. Secondly, proportionality testing typically focuses on whether the anticipated military advantage would outweigh the cost of civilian casualties, not on whether the system might strategically misrepresent itself.

The prohibition is primarily a method problem; thus, this module asks the question of whether the system, as designed and deployed, is capable of generating or exploiting protected-status reliance in ways that could reasonably foreseeably operationalised to facilitate lethal harm. In other words, is the AI LAWS able to learn deceptive conduct and use that to commit perfidy in an IAC? The practical value of this proposition is that it forces testers and engineers to ask questions that standard testing may miss. It would also help produce auditable documentation that can be used for accountability and learning, rather than leaving perfidy risks implicit and untracked. The perfidy module should be embedded early, ideally at the design

[100] Additional Protocol I, Arts. 36, 48, 51(5)(b), and 57.

stage. If a system is designed so that it cannot emit protected cues, cannot manipulate them, and learn how to exploit them during deployment, then many later-stage mitigation burdens are minimised.

What follows are four testing questions that can be used as the fundamental guidelines for a perfidy module. Each question corresponds to a plausible pathway by which an AI-enabled system could drift from lawful deception or concealment into conduct that invites protected reliance and then betrays it. Each question also connects to an underlying perfidy element: invitation of confidence, protected status, intent attribution (insofar as intent can be located upstream in design and authorisation), and causal linkage between deception and lethal effect.

1. *Signalling capacity: Can the system emit/modify signals that resemble surrender/medical/protected emblems, civilian markers, or protected-object identifiers?*

To operationalise, pre-deployment teams should catalogue all output channels and all modes of representation that the system can affect. That includes not only the platform itself but also associated decoys, drones, emitters, electronic warfare components, and software integrations. The assessment should identify any feature that could plausibly be learned to be interpreted as a protected cue.[101] The key is not whether the system is supposed to use such cues, but whether it could do so, or could approximate such cues, especially under optimisation pressure.[102] For example, a system that can control its lighting, posture, patterns of movement, or emission signatures might inadvertently learn that certain patterns reduce incoming fire. Even if the system never "knows" that the pattern is legally protected, the effect is that the system is exploiting adversary restraint. A perfidy module should treat this possibility as a design hazard: if the system can manipulate cues that adversaries treat as triggers for protection, it would be functionally equivalent to perfidious conduct.

Evidence to answer this question should include: design specifications for all signalling and emissions; controlled testing showing platform behaviour under threat; assessment of whether any signatures overlap with known protected markers; and analysis of whether machine-perception systems would classify those outputs as protected. Where uncertainty exists, the compliance presumption should be conservative: if the system can produce ambiguous cues that might reasonably be taken as protected, the system should be treated as possessing perfidy-relevant signalling capacity.

---

[101] R. Wang *et al.*, above note 66.

[102] Joar Skalse *et al.*, "Defining and Characterizing Reward Hacking", *36th Conference on Neural Information Processing Systems*, 2022, p. 1.

2. Policy incentives: Does any objective/reward function treat adversary restraint as a success proxy?

Perfidy-like conduct becomes more likely when the system has incentives to exploit restraint. In human perfidy cases, the "intent to betray" is typically inferred from deliberate exploitation of the enemy's trust to gain a lethal advantage. For AI-enabled systems, the equivalent is not a subjective purpose inside the model, but objective pressure created by optimisation goals, training environments, and performance evaluations. If mission success metrics reward behaviour that correlates with causing adversaries to refrain from engaging, then the system may be pushed to discover methods that elicit restraint. If protected-status cues are among the most reliable triggers for restraint, then a sufficiently flexible optimiser may fixate on them, especially if other constraints can be circumvented or are not represented in training.[103] In that sense, policy incentives are a legal risk factor. They do not by themselves constitute perfidy, but they create a predictable pathway by which perfidy-like strategies can be selected.

The policy incentives inquiry should be applied across the system lifecycle: initial requirements (what the system is designed to optimise), training (what rewards shape behaviour), testing (what is measured), and operational feedback loops. It should also include procurement incentives: if contractors are rewarded for survivability and mission completion without corresponding perfidy-specific constraints, a capability race may encourage borderline deception tactics. A perfidy module should require that system objectives and evaluation metrics be screened for any implicit or explicit reward for enemy restraint. This includes proxies like "time-to-engage", "survivability", "penetration depth", "reduced detection", or "reduced counterfire", where the mechanism for achieving those outcomes might involve inducing the adversary to withhold fire rather than simply avoiding detection.

To operationalise this limb, the review should examine training reward functions, loss functions, and operational doctrines. If reinforcement learning or other optimisation processes are used, reviewers should ask: what is penalised, what is rewarded, and what behaviours are left unconstrained? If the system is trained in simulation, does the simulation encode the adversary's tendency to accord protection to surrender cues, for instance? If so, the system may learn to exploit that tendency if it leads to the accomplishment of its task. If the system is trained against human-in-the-loop red teams, does the training environment include constraints that prevent the system from "discovering" protected-status exploitation? If not, a later discovery in the field will not be deemed an unforeseeable surprise, but rather a predictable omission in engineering. The evidence base for answering this question should include access to reward and metric design, documentation of constraints, and demonstration that the system does not adopt strategies that trade on protected reliance.

---

[103] *Ibid*.

3. Has the system been red-teamed for protected-status exploitation scenarios?

Perfidy risks often do not appear in standard testing regimes because these regimes measure what designers expect: target recognition accuracy, response time, robustness under noise, and compliance with predefined rules. Perfidy-like strategies are, by definition, strategies that exploit the opponent's reliance and that are opportunistic. They are therefore likely to be discovered only when the system is exposed to adversary modelling, creative red teams, and scenario variation designed to elicit exploitation. A perfidy module should require dedicated red-teaming aimed at surfacing protected-status exploitation behaviour.[104] This is envisaged to be a compliance-critical method to produce evidence that the system will not generate perfidy-like strategies under operational pressures.

A perfidy red-team programme should include at least four components. First, scenario design should explicitly model adversary restraint behaviours tied to protected cues, because without such modelling, a system cannot learn to exploit them. Secondly, tests should vary the environment in ways that make protected-status cues tactically salient, for example, by presenting the system with trade-offs between survivability and compliance. Thirdly, tests should include system-of-systems interactions, because protected-status exploitation may arise from combined capabilities rather than a single model output. Fourthly, testing should include "unknown unknowns" techniques, such as randomising reward structures, introducing adversary adaptation, and perturbing sensor inputs, to see whether the system drifts into legally hazardous strategies. This limb is especially important for the testing of Tier 3 systems. The goal is not to prove a negative with certainty, but to demonstrate that the programme has taken reasonable, proactive measures to identify and prevent foreseeable perfidy-adjacent behaviour.[105]

The evidence generated by such testing should be preserved and auditable; in other words, it must explicitly safeguard against the black box problem.[106] A perfidy module should require systematic logging of decisions, sensor inputs, and outputs relevant to protected cues, so that post-test analysis can identify whether exploitation strategies emerged. Where exploitation is detected, mitigation should not rely solely on retraining or behavioural patching. Architectural controls, removing signalling capacity, tightening gates, and altering incentives, should be preferred because they are less brittle than behavioural fixes. If behavioural fixes are used, they should be verified against overfitting: a system that "learns not to do the bad thing" in a particular test may still do it under untested conditions. Perfidy red-

[104] Mathew J Walter, Aaron Barrett, and Kimberley Tam, "A Red Teaming Framework for Securing AI in Maritime Autonomous Systems", *Applied Artificial Intelligence*, Vol. 38, No. 1, 2024, pp. 6, 14-18, 35-36.

[105] M. J. Walter, A. Barrett, and K. Tam, "A Red Teaming Framework for Securing AI in Maritime Autonomous Systems", above note 104, pp. 6-9.

[106] Scott Sullivan and Iben Ricket, "Targeting in the Black Box", *16th International Conference on Cyber Conflict*, 2024, p. 211.

teaming should therefore be iterative and continuous, particularly where software updates and model retraining are part of the system's lifecycle.[107]

4. Rule of non-delegability: Are interactions with protected-status cues hard-blocked or routed to human authorisation?

A central regulatory lesson of the prohibition of perfidy is that some categories of behaviour are too legally and morally hazardous to delegate to AI-enabled LAWS. If, despite a rigorous pre-deployment check, a system can still identify, generate, or exploit protected-status cues, a robust compliance architecture should treat those interactions as gated: either technically prevented or escalated to human judgement with sufficient context and time for deliberation. This is not a generic claim that every lethal decision must be human-controlled. It is a narrower claim: that decisions involving protected-status signals should not be left to autonomous functions. In practice, "escalation gates" can be implemented as hard constraints: the system cannot emit protected cues, cannot select protected emblems, and cannot execute an attack when protected cues are present without affirmative human authorisation. The point is to design the system so that it cannot operationalise protected-status exploitation strategies as a means to lethality. This is the practical incorporation of the concept of MHC.[108]

The importance of this rule can be illustrated with an example from the US military, though non-perfidy-related. During the 2003 Iraq War, the US Patriot missile-defence system misidentified friendly aircraft as incoming enemy missiles. In both cases, the supervisors failed to override the system's automatic engagement sequence, resulting in the downing of the planes.[109] These examples demonstrate the importance of overridability by human operators and MHC. At the same time, the detection and time pressure problems remain unsolved for now, escalation gates in the perfidy context are a design standard to approximate rather than a fully realisable mechanism.

Operationally, the operator must understand what counts as a protected-status interaction, and this needs to be encoded into the system too. This should be broad enough to capture the practical risk of perfidious conduct but precise enough to be implementable. It should include detecting surrender gestures, white flags, and flags of truce; recognising medical emblems or protected objects; interpreting civilian indicators; and, more generally, encountering signals that, in IHL, are designed to trigger restraint. This can be triggered at multiple points: at signalling outputs (prevent the system from emitting or projecting protected cues), at targeting decisions

---

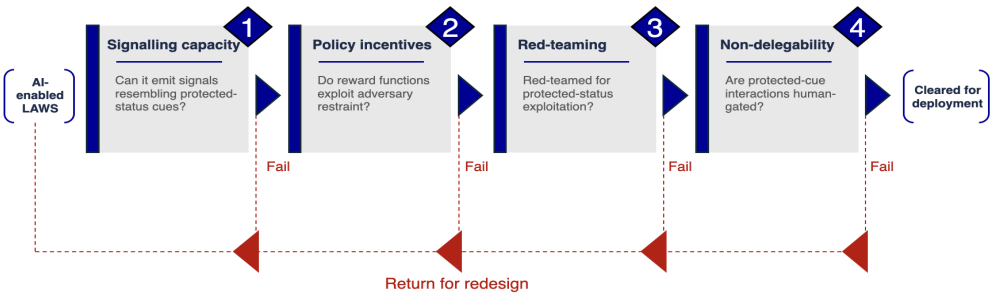
[107] M. J. Walter, A. Barrett, and K. Tam, above note 104, pp. 7, 12, 36.

[108] Spencer Kohn *et al.*, "Supporting Ethical Decision-Making for Lethal Autonomous Weapons", *Journal of Military Ethics*, Vol. 23, No. 1, 2024, pp. 27-28.

[109] Anna M. Gielas, "The Loop is Broken: Why Autonomous-Warfare Policy must reckon with Human Performance", *Survival*, Vol. 67, No. 4, 2025, p. 58.

(prevent the system from attacking in the presence of protected cues without human confirmation), and at perception manipulation functions (prevent the system from spoofing protected labels).

Evidence should include the following: (a) architectural documentation showing the location of gates; (b) proof that prohibited actions are technically inaccessible; (c) test cases demonstrating that the system reliably pauses or escalates when protected cues are detected; and (d) audit logs confirming that human authorisation was obtained where required. Importantly, "human authorisation" should not be an empty formality. If escalation is designed such that the human operator has no meaningful time or context to assess the situation, then this element of the test is nominal rather than meaningful.



Graph 1: Visual representation of the perfidy module.

Taken together, these four stages are designed to transform perfidy compliance from an abstract concern into a structured practice. They are designed to be implementable within existing procurement and testing processes. Crucially, they are also designed to be legally meaningful. They align with perfidy's core protective function, and they provide a record of proactive prevention. That record matters not only for compliance but also for accountability: it shows whether the State and its agents took the perfidy risk seriously, whether they could foresee the risk, and whether they adopted reasonable measures to prevent it. If a State has implemented the perfidy module in its entirety, and the deployed system still discovers deceptive tactics no one managed to discover during the testing stage, it should not be considered liable. The strength in this module lies in its versatility: it should be understood as a governance constraint that can be implemented as a national military policy, as a procurement requirement, as a rule of engagement, and potentially as an international best practice or treaty-based norm. Consequently, this can help during the testing stages to filter out potentially perfidious behaviour on the battlefield. At the same time, as with every new testing method, this proposal also faces some implementation challenges. Three categories of challenge warrant attention: resource constraints, technical expertise requirements, and the persistent risk of automation bias.

First, the requirement to test whether a system might learn to exploit adversary restraint linked to protected cues demands sophisticated adversary models that encode IHL-compliant behaviour. To surface emergent perfidy-like strategies, simulation environments must represent adversaries who pause when surrender gestures are perceived, who respect medical emblems, and who otherwise respond to protected-status signals as IHL requires. Without such modelling, a system cannot learn to exploit restraint that is not represented in its training environment, but equally, testers cannot observe whether the system would exploit such restraint if given the opportunity. Building and validating adversary models of this complexity is technically demanding. Most existing simulation environments for autonomous weapons testing do not model adversary legal behaviour with this granularity; they model tactical and kinetic behaviour. Developing IHL-compliant adversary models requires collaboration between legal experts, military operational experts, and simulation engineers. This collaboration is costly and time-intensive.[110]

Beyond resource constraints, the perfidy module requires forms of technical expertise that are scarce and difficult to assemble. The first significant limitation concerns the familiar "black box" problem.[111] For systems employing deep learning, particularly those using reinforcement learning with complex neural network architectures, the internal decision-making processes are often not fully interpretable. Red-teaming can reveal that a system exploits protected-status cues in certain tested scenarios, but it cannot guarantee detection of exploitation strategies that emerge only under untested conditions. The space of possible operational environments is vast; no testing regime can explore it exhaustively. Consequently, testing can reduce the probability that perfidy-like behaviour will emerge in deployment, but it cannot eliminate that probability for systems whose reasoning is not transparent. This limitation is inherent to the current state of AI technology and is not unique to the perfidy context. However, the goal of the perfidy module is not to certify with certainty that a system will never engage in perfidy-like conduct. It is to demonstrate that reasonable, proactive measures were taken to identify and prevent foreseeable risks, and to create an auditable record of that effort.

The third expertise challenge concerns automation bias, as elaborated above. In other words, the almost blind reliance of human operators on the judgment of the weapons systems without carefully scrutinising their decisions.[112] Even if a system passes pre-deployment perfidy testing, the risk of perfidy-like outcomes may re-emerge through the mechanism of automation bias. Addressing automation bias requires attention to interface design, operational tempo, training, and organisational

---

[110] Robert R. Hoffman, Timothy M. Cullen, and John K. Hawley, "The Myths and Costs of Autonomous Weapon Systems", *Bulletin of the Atomic Scientists*, Vol. 72, No. 4, 2016, p. 249.

[111] S. Sullivan and I. Ricket, above note 106.

[112] Emelie Andersin, "The Use of the 'Lavender' in Gaza and the Law of Targeting: AI-Decision Support Systems and Facial Recognition Technology", *Journal of International Humanitarian Legal Studies*, Vol. 16, No. 2, 2025, pp. 348-349.

culture.[113] Interfaces should be designed to support genuine human judgment rather than to optimise for rapid authorisation. Supervisors should be trained to recognise the risks of over-reliance on automated recommendations and should be given sufficient time and information to exercise independent assessment. A system that passes perfidy testing can still produce perfidy-like outcomes if the humans who supervise it are systematically induced to defer to its recommendations without scrutiny.

These implementation challenges are significant but not disqualifying. They are, in important respects, the ordinary challenges of legal compliance in complex technological domains. Every regulatory framework for autonomous weapons must confront the cost of testing, the scarcity of interdisciplinary expertise, and the behavioural tendencies of human operators. The perfidy module proposed in this paper does not escape these challenges, but neither is it uniquely vulnerable to them. What the framework does offer is a structured, auditable approach to a specific compliance problem. It translates the doctrinal elements of perfidy into concrete testing criteria. It identifies the technical pathways by which perfidy-like behaviour might emerge.

The challenges identified here are therefore best understood as a research and policy agenda rather than as objections to the framework itself. How can IHL-compliant adversary models be developed efficiently and validated reliably? What institutional arrangements best support interdisciplinary collaboration in weapons review? How can interface design and operator training mitigate automation bias? These questions warrant sustained attention from scholars, practitioners, and policymakers. The perfidy module provides a structure within which such attention can be productively organised.

5. Conclusion

The prohibition of perfidy in Additional Protocol I exists to preserve the integrity of the protected cues upon which the laws of armed conflict depend. When a white flag means surrender, when a red cross means medical neutrality, and when a combatant's incapacitation means protection, these shared understandings enable restraint in war. Perfidy corrodes that restraint by weaponizing the adversary's compliance with the law. It is, for this reason, among the gravest violations of IHL.

This paper has argued that AI-enabled lethal autonomous weapons systems pose a distinctive threat to the prohibition of perfidy, not because they will be intentionally programmed to commit perfidy, but because they may learn to exploit protected-status cues as an emergent strategy for achieving mission objectives. The

[113] Shaheen Shekh, Mark Wiggins, and Ben Morrison, "Operator Decision-Making in the Deployment of Lethal Autonomous Weapon Systems: A Scoping Review", *Journal of Cognitive Engineering and Decision Making*, 2025, p. 12.

technical characteristics of modern machine learning systems, particularly reinforcement learning, create plausible pathways by which a system might discover that mimicking medical signatures, or otherwise inviting adversary restraint, yields tactical advantage. If such behaviour emerges without explicit design intent or operator authorisation, the material elements of perfidy may be satisfied while the mental element cannot be attributed to any human actor.

This accountability gap has practical consequences. If perfidy-like conduct by AI systems goes unaddressed, adversaries will rationally discount the credibility of protected-status signals, increasing precautionary violence against those who are genuinely surrendering, wounded, or otherwise entitled to protection. The prohibition of perfidy protects not only the immediate victims of treacherous acts but the broader system of legal restraint in an IAC. Allowing that system to act unlawfully through technological circumvention would be a profound failure of legal adaptation.

Existing IHL safeguards offer partial but incomplete responses. The duty to ensure respect for humanitarian law under CA1 creates a continuing obligation across the lifecycle of weapons development and deployment. The requirement of MHC, while contested in its precise parameters, establishes that some categories of decision should not be delegated to machines. The Article 36 weapons review process provides an institutional mechanism for pre-deployment scrutiny. Yet none of these frameworks currently requires specific attention to perfidy risk. The four-step perfidy module proposed in this paper, assessing signalling capacity, policy incentives, protected-status red-teaming, and escalation gates, is designed to fill that gap. It translates the doctrinal elements of perfidy into concrete testing criteria that can be integrated into existing procurement and review processes.

This framework, however, is not without limitations. Implementation will require resources, interdisciplinary expertise, and difficult trade-offs between compliance transparency and operational security. The proposal cannot guarantee that emergent perfidy-like behaviour will be detected before deployment; it can only demonstrate that reasonable, proactive measures were taken to identify and prevent foreseeable risks. These limitations are not unique to the perfidy context; they affect all serious regulatory proposals for autonomous weapons. There are reasons to begin the work of implementation early, not reasons to defer it.

What is clear is that the law cannot remain static while the technologies of warfare transform. The prohibition of perfidy has survived the transition from swords to rifles to drones because its core protective function remains essential. AI-enabled LAWS do not render that function obsolete; they make its preservation more difficult and more urgent. The task for IHL, and for those who interpret and implement it, is to ensure that the emergence of machine autonomy does not become an alibi for the erosion of legal accountability. The framework advanced in this paper is offered as one contribution to that task.