INDIAN JOURNAL OF LEGAL REVIEW [IJLR – IF SCORE – 7.58]

VOLUME 6 AND ISSUE 2 OF 2026

APIS – 3920 - 0001 *(and)* ISSN - 2583-2344

Published by

Institute of Legal Education

https://iledu.in

THE BLACK BOX PROBLEM IN AI AND ITS IMPACT ON ATTRIBUTION IN CYBER CONFLICTS: INTERNATIONAL LAW'S RESPONSE TO UNTRACEABLE DECISION-MAKING

AUTHOR – ALLEN BENNY, BBA LLB, SCHOOL *OF* LAW, CHRIST (DEEMED TO BE UNIVERSITY), BENGALURU

BEST CITATION – ALLEN BENNY, THE BLACK BOX PROBLEM IN AI AND ITS IMPACT ON ATTRIBUTION IN CYBER CONFLICTS: INTERNATIONAL LAW'S RESPONSE TO UNTRACEABLE DECISION-MAKING, *INDIAN JOURNAL OF LEGAL REVIEW (IJLR),* 6 (2) OF 2026, PG. 664-671, APIS – 3920 – 0001 & ISSN - 2583-2344

Abstract

The rapid militarization of artificial intelligence in cyber operations highlights a growing concern in international law. The "black box" nature of deep-learning systems means that their autonomous decision-making processes are untraceable, even to their creators. This type of technological opacity undermines attribution, a doctrine that is concerned with proof of intent, control (effective, overall, or in a causal sense), and the law of international responsibility, including the ARSIWA, the Tallinn Manual 2.0, and key case law, including the Nicaragua and Tadić cases. When fully autonomous AI systems behave erratically and unpredictably, the law of international responsibility and international humanitarian law (IHL) are left with critical gaps, flouting the principles of sovereignty, distinction, proportionality, and the law of armed conflict (meaningful human control).

Employing a doctrinal, comparative, and interdisciplinary methodology, this paper analyses how algorithmic opacity disrupts legal attribution in cyber conflicts and evaluates whether Explainable AI (XAI) techniques can restore transparency and traceability. While XAI cannot eliminate the black box entirely, it offers practical audit trails and evidentiary support for Article 36 weapons reviews and post-incident investigations. The findings demonstrate that current international legal instruments remain ill-equipped for non-human agency. The paper concludes that meaningful reform—incorporating a legal duty of explainability, constructive control standards, and institutional oversight mechanisms—is essential to preserve accountability in the age of autonomous cyber warfare.

Keywords: Artificial Intelligence, Black Box Problem, State Responsibility, Cyber Warfare, Attribution, Explainable AI, Constructive Control.

I. Introduction

In June 2025, the state-sponsored group APT28, associated with the Russian government, used PROMPTSTEAL malware for the first time as a self-learning, state-sponsored AI malware for cyber operations against Ukrainian targets, as it incorporates real-time autonomous queries to large language models to create and execute commands.[1759] This constitutes one of the first instances of self-adaptive AI being used in state-sponsored cyber operations. The use of self-learning malware in cyber operations illustrates a rapid AI militarization and a significant gap in international law: the "black box" issue. The "black box" problem describes the problem of a deep-learning system that arrives at a particular decision through a series of processes that, as a result of the deep-learning system, remain unknown to the

[1759] Google Threat Intel. Grp., Advances in Threat Actor Usage of AI Tools (Nov. 5, 2025), https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools (reporting APT28's deployment of PROMPTSTEAL malware that queries LLMs in live operations against Ukraine).

INDIAN JOURNAL OF LEGAL REVIEW [IJLR – IF SCORE – 7.58]
VOLUME 6 AND ISSUE 2 OF 2026
APIS – 3920 - 0001 *(and)* ISSN - 2583-2344
Published by
Institute of Legal Education
https://iledu.in

system's operators, leading to a state of responsibility, purpose, and control that cannot be attributed to any particular person.[1760]

Attribution is the process of determining state responsibility and liability under the law of international humanitarian law (IHL). When autonomous AI systems breach and trespass by manipulating or disrupting systems without any apparent human involvement, the traditional legal doctrines fall apart. The Articles on Responsibility of States for Internationally Wrongful Acts (ARSIWA)[1761] and the control tests developed in *Military and Paramilitary Activities* in and Against *Nicaragua* (Nicar. v. U.S.)[1762] and *Prosecutor v. Tadić*[1763] presuppose identifiable human agency and intent. The Tallinn Manual 2.0 similarly assumes rational human decision-making.[1764] The absence of human control leads to unaccountability in the law, which undermines the principles of sovereignty, distinction, proportionality, and the meaningful human control.

This leads to the main research question of the paper: to what extent can current international legal systems ascribe responsibility to opaque autonomous AI systems during cyber warfare? The analysis is guided by three subsidiary questions: (1) To what extent does the black box phenomenon frustrate traditional attribution doctrines? (2) To what extent can artificial intelligence–driven warfare be regulated by the United Nations Charter, ARSIWA, IHL, and the Tallinn Manual 2.0? and (3) To what extent can Explainable AI (XAI) provide a means to restore transparency and accountability?

The thesis is straightforward: current international law, grounded in anthropocentric assumptions of intent and control, is ill-equipped to address non-human agency in cyber warfare. While XAI techniques offer partial technical mitigation through audit trails and interpretability, meaningful reform—introducing a legal duty of explainability, a "constructive control" standard, and institutional oversight—is essential to preserve the rule of law.

This study employs a doctrinal, comparative, and interdisciplinary methodology, analysing primary legal sources alongside technological literature on deep learning and XAI. The paper proceeds as follows: Part II builds the conceptual groundwork of attribution and AI opacity; Part III analyses the manner in which the black box in practice undermines accountability; Part IV discusses the shortcomings of current possibilistic frameworks; Part V posits XAI as a possibilistic bridge; Part VI presents specific reform suggestions; and Part VII will end with the necessity of the legal frameworks evolution.

II. Foundations: Attribution Theory and AI Opacity

2.1 Attribution and State Responsibility

Attribution is the doctrinal linchpin of state responsibility in international law. Under the Articles on Responsibility of States for Internationally Wrongful Acts (ARSIWA), a state is responsible for conduct attributable to it, including acts of its organs or entities exercising governmental authority (Articles 4–8) or conduct directed or controlled by the state (Article 8).[1765] The International Law Commission's commentary emphasises that attribution requires a sufficiently close link between the act and the state.[1766]

Two landmark decisions have shaped the required degree of control. In *Military and Paramilitary Activities in and Against Nicaragua*, the International Court of Justice applied the stringent "*effective control*" test, holding that the United States could be responsible for Contra operations only if it had directed or enforced the

[1760] Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information 3–8 (2015).
[1761] Articles on Responsibility of States for Internationally Wrongful Acts, G.A. Res. 56/83, Annex, U.N. Doc. A/RES/56/83 (Dec. 12, 2001) [hereinafter ARSIWA].
[1762] Military and Paramilitary Activities in and Against Nicaragua (Nicar. v. U.S.), Judgment, 1986 I.C.J. 14, ¶¶ 109–15 (June 27).
[1763] Prosecutor v. Tadić, Case No. IT-94-1-A, Judgment, ¶¶ 98–145 (Int'l Crim. Trib. for the Former Yugoslavia July 15, 1999).
[1764] Michael N. Schmitt, Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations 29–35, 84–88 (2d ed. 2017).
[1765] ARSIWA, supra note 3, arts. 4–8.
[1766] Int'l L. Comm'n, Draft Articles on Responsibility of States for Internationally Wrongful Acts, with Commentaries, 2 Y.B. Int'l L. Comm'n 31, 38–45 (2001).

INDIAN JOURNAL OF LEGAL REVIEW [IJLR – IF SCORE – 7.58]

VOLUME 6 AND ISSUE 2 OF 2026

APIS – 3920 - 0001 *(and)* ISSN - 2583-2344

Published by

Institute of Legal Education

https://iledu.in

specific wrongful conduct.[1767] The International Criminal Tribunal for the Former Yugoslavia later adopted a broader "overall control" standard in *Prosecutor v. Tadić*, finding that general control over a group's activities suffices for attribution of international crimes.[1768] Both tests, however, presuppose identifiable human agency and intentional direction.

The Tallinn Manual 2.0 extends these principles to cyberspace. Rule 6 states that cyber operations by state organs or agents are attributable to the state, while Rule 15 addresses proxy actors acting on behalf of a state.[1769] Yet the Manual explicitly assumes human decision-making and leaves unresolved how attribution operates when the "agent" is an autonomous algorithm.[1770]

2.2 The Black Box Problem – Technical Primer

Issues regarding "black box" problems often occur from the structure present within today's modern deep-learning systems. Rather than being rule-based and algorithmic, deep neural networks involve several hidden layers with millions to billions of tunable parameters responsible for changes during training and back-propagation.[1771] After training, the model will produce inputs and outputs, as a result of a series of non-linear activations and weighted connections. Unfortunately, no one can explain the individual contributions from each of the connections.[1772] This opacity is not a flaw but an inherent feature: the system's decision logic emerges from statistical correlations rather than explicit rules.[1773]

Modern deep learning systems are used to perform cyber operations, specifically for autonomous intrusion detection, target selection, adaptive malware, etc. An example of this would be an AI driven cyber defense system, a system which can detect and respond to intrusions autonomously. Because of this, the system may respond to an intrusion by launching one of the learned countermeasures, even when the system itself may not understand the reasons for doing so. Ultimately, all systems will lack an explainable audit trail and evidence should be interpreted as a system failing. For example, forensic analysis, whether it be predicting a cyber attack, or attempting to understand how or why the system acted as it did will be made impossible.[1774]

2.3 The Accountability Gap

This technological opacity collides directly with the legal requirements of intent (*mens rea*) and control. As Sandra Wachter and Luciano Floridi have demonstrated, when an AI system produces harm that neither its operators nor developers can explain, traditional notions of foreseeability and causation dissolve.[1775] The accountability gap that emerges is not merely evidentiary; it is structural. States can neither prove compliance with international humanitarian law nor be held meaningfully responsible under ARSIWA when the "actor" is an inscrutable algorithm.[1776] The foundational assumption of human agency that underpins both state responsibility and IHL is therefore rendered untenable.

III. The Collision: How AI Opacity Erodes Attribution in Cyber Conflicts

3.1 Typology of AI-Enabled Cyber Operations

With AI, cyber operations can be modified or improved in three different ways, while attribution can be made even more difficult with each adaptation or enhancement. For example, concerning state sponsored attacks,

[1767] Military and Paramilitary Activities in and Against Nicaragua (Nicar. v. U.S.), Judgment, 1986 I.C.J. 14, ¶¶ 109–15 (June 27).
[1768] Prosecutor v. Tadić, Case No. IT-94-1-A, Judgment, ¶¶ 98–145 (Int'l Crim. Trib. for the Former Yugoslavia July 15, 1999).
[1769] Michael N. Schmitt, Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations 29–35, 84–88 (2d ed. 2017).
[1770] Id. at 30–32.
[1771] Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information 3–8 (2015).
[1772] Jenna Burrell, How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms, 3 Big Data & Soc'y 1, 4–6 (2016).
[1773] See Pasquale, supra note 13, at 9–12.
[1774] Google Threat Intel. Grp., Advances in Threat Actor Usage of AI Tools (Nov. 5, 2025), https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools.
[1775] Sandra Wachter & Luciano Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, 7 Int'l Data Privacy L. 76, 82–85 (2017).
[1776] See Rebecca Crootof, War, Responsibility, and Killer Robots, 40 N.C. J. Int'l L. 909, 932–35 (2015).

INDIAN JOURNAL OF LEGAL REVIEW [IJLR – IF SCORE – 7.58]

VOLUME 6 AND ISSUE 2 OF 2026

APIS – 3920 - 0001 *(and)* ISSN - 2583-2344

Published by

Institute of Legal Education

https://iledu.in

AI can be used to augment attacks in a human-directed manner. Take the example of the 2020 SolarWinds Supply Chain Attack. The attack involved the use of machine learning to navigate attacks while avoiding the detection of human operatives who completed the extraction.[1777] Second, fully autonomous operations allow systems to identify targets, launch intrusions, and adapt in real time without further human input. In 2025, Russian operation APT28 documented malware PROMPTSTEAL, which was able to autonomously summon an evasion code generation large language model in the midst of active combat in Ukraine.[1778] Operations targeted with algorithmic misinformation can use deepfakes, bot networks, public opinion manipulation, and even an unneutralized state of conflict. The lack of transparency with the use of AI unchecked can produce an effect comparable to the classical state operations of way without crossing state boundaries.[1779]

3.2 Core Analytical Challenges

Each typology collides with the legal requirements of intent and control. Consider a running hypothetical: State A deploys an autonomous AI cyber-defense system ("ShieldAI") along its border. When ShieldAI detects anomalous traffic it classifies as hostile, it autonomously retaliates by disrupting a civilian hospital's power grid in State B. No human operator authorised or even reviewed the specific target.[1780]

Under the *Nicaragua* effective-control test and the *Tadić* overall-control test, attribution fails because there is no evidence of state direction over the specific act.[1781] The AI's decision emerged from millions of learned parameters, not from any human command. Mens rea evaporates: the deploying state cannot prove it lacked intent, nor can the victim state prove the state possessed it.[1782]

Evidentiary barriers compound the problem. Traditional digital forensics relies on log files, IP addresses, and command-and-control infrastructure. ShieldAI's deep neural network leaves no intelligible trace of why it selected the hospital network; post-incident reconstruction is mathematically impossible.[1783] Even if investigators recover the model weights, they reveal only statistical correlations, not legal causation.[1784]

Dual attribution further complicates responsibility. Private contractors often train and sell these models. If a developer in a third country supplied the flawed training data, is liability shared with State A? ARSIWA Article 8 requires state "direction or control," yet the developer exercised no operational control after deployment.[1785] Product-liability analogies from domestic law offer no clear international solution.[1786]

3.3 Strategic Ambiguity and Plausible Deniability

States can intentionally use the black box. Through the use of highly autonomous systems, a state can achieve plausible deniability that traditional proxies would not match, as states can invoke "technological malfunction" or "unintended emergence".[1787] In the ShieldAI scenario, State A would be correct to state that there was no order given for the attack on the

[1777] Microsoft Threat Intelligence Ctr., SolarWinds Supply Chain Compromise (Dec. 2020), https://www.microsoft.com/en-us/security/blog/2020/12/13/solarwinds-supply-chain-compromise (describing AI-assisted network mapping).

[1778] Google Threat Intel. Grp., Advances in Threat Actor Usage of AI Tools (Nov. 5, 2025), https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools.

[1779] See Duncan B. Hollis, The Role of International Law in Cyber Operations, 111 Am. J. Int'l L. 291, 312–15 (2017).

[1780] This hypothetical is constructed for analytical purposes and draws on documented patterns in autonomous cyber-defense systems.

[1781] Military and Paramilitary Activities in and Against Nicaragua (Nicar. v. U.S.), Judgment, 1986 I.C.J. 14, ¶¶ 109–15 (June 27); Prosecutor v. Tadić, Case No. IT-94-1-A, Judgment, ¶¶ 98–145 (Int'l Crim. Trib. for the Former Yugoslavia July 15, 1999).

[1782] See Rebecca Crootof, War, Responsibility, and Killer Robots, 40 N.C. J. Int'l L. 909, 932–35 (2015).

[1783] Jenna Burrell, How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms, 3 Big Data & Soc'y 1, 4–6 (2016).

[1784] Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information 9–12 (2015).

[1785] Articles on Responsibility of States for Internationally Wrongful Acts, G.A. Res. 56/83, Annex, art. 8, U.N. Doc. A/RES/56/83 (Dec. 12, 2001) [hereinafter ARSIWA].

[1786] *See* generally Kenneth Anderson & Matthew C. Waxman, Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can, Hoover Inst. (2013).

[1787] Michael N. Schmitt, Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations 30–32 (2d ed. 2017).

INDIAN JOURNAL OF LEGAL REVIEW [IJLR – IF SCORE – 7.58]
VOLUME 6 AND ISSUE 2 OF 2026
APIS – 3920 - 0001 *(and)* ISSN - 2583-2344
Published by
Institute of Legal Education
https://iledu.in

hospital, as there would be no order to attack, and still be able to reap the benefits of the disruption. Such strategic ambiguity diminishes existing deterrence, and the entire attribution system.[1788]

3.4 Accountability Gap in Practice

The cumulative result is a structural accountability gap. Neither ARSIWA nor the Tallinn Manual 2.0 contemplates non-human agency.[1789] When ShieldAI causes civilian harm, State B cannot satisfy the burden of proof before the International Court of Justice or trigger lawful countermeasures under Article 51 of the UN Charter.[1790] International humanitarian law principles—distinction, proportionality, and precaution—become unenforceable because no human actor can demonstrate compliance or violation.[1791] The gap is not merely procedural; it erodes the normative foundation of the international legal order itself.

IV. Limits of Existing International Legal Frameworks

The existing international legal architecture was designed for human actors, not opaque algorithms. Misalignment becomes apparent when tested against the ShieldAI hypothetical from Section III.

4.1 United Nations Charter

The UN Charter's Articles 2(4) and 51 forbid the use of force, although self-defense against an "*armed attack*" is permitted.[1792] The Tallinn Manual 2.0 views major cyber operations with physical or functional damage as possible armed attacks.[1793] However, with ShieldAI, if an operation autonomously neutralizes a civilian hospital's power grid, one cannot determine the operation as a state-sanctioned armed attack or just an unpredictable algorithmic decision. The intent-based threshold of the Charter cannot be used here.[1794]

4.2 ARSIWA and the Tallinn Manual 2.0

For attribution purposes, the ARSIWA Articles 4-8 and 11 require characteristics of direction or control.[1795] The Tallinn Manual 2.0 (Rules 6 and 15) applies this reasoning to cyberspace but presumes, rather pointedly, human involvement.[1796] In the case of ShieldAI, neither effective control nor overall control can be ascertained because the determining action came from the model's opaque parameters and not from any state command. The Manuals and Articles therefore offer no workable solution when the "*agent*" is an autonomous system.[1797]

4.3 International Humanitarian Law

The Geneva Conventions assume that ethics and human judgment are at play concerning the application of the principles of distinction, proportionality, and precaution.[1798] An opaque AI system, however, cannot and does not distinguish between military and civilian targets, and leaving the reasoning to the deploying state is a failure of the precaution obligation. Article 36 of Additional Protocol I requires States to review the legality of new weapons to ensure compliance with the principles of distinction, proportionality, and precaution. [1799] The result is a complete failure of the precaution obligation.

The absence of accountability in autonomous cyber warfare is an obligation to review the principles of ethics in warfare. The principles of ethics in warfare are absent in autonomous cyber warfare, and in traditional warfare the principles of ethics in warfare are robust. But

[1788] *See* Heather M. Roff, The Strategic Implications of Autonomous Weapons, 25 Ethics & Int'l Aff. 1, 12–15 (2019).
[1789] Schmitt, supra note 29, at 84–88.
[1790] U.N. Charter art. 51.
[1791] Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), art. 51, June 8, 1977, 1125 U.N.T.S. 3 (distinction and proportionality).
[1792] U.N. Charter art. 2, ¶ 4; id. art. 51.
[1793] Michael N. Schmitt, Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations 126–30 (2d ed. 2017).
[1794] See Duncan B. Hollis, The Role of International Law in Cyber Operations, 111 Am. J. Int'l L. 291, 312–15 (2017).
[1795] Articles on Responsibility of States for Internationally Wrongful Acts, G.A. Res. 56/83, Annex, arts. 4–8, 11, U.N. Doc. A/RES/56/83 (Dec. 12, 2001) [hereinafter ARSIWA].
[1796] Schmitt, supra note 35, at 29–35, 84–88 (Rules 6 & 15).
[1797] Id. at 30–32.
[1798] Geneva Convention Relative to the Protection of Civilian Persons in Time of War, art. 27, Aug. 12, 1949, 75 U.N.T.S. 287.
[1799] Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), art. 36, June 8, 1977, 1125 U.N.T.S. 3.

INDIAN JOURNAL OF LEGAL REVIEW [IJLR – IF SCORE – 7.58]

VOLUME 6 AND ISSUE 2 OF 2026

APIS – 3920 - 0001 *(and)* ISSN - 2583-2344

Published by

Institute of Legal Education

https://iledu.in

with autonomous cyber warfare, traditional principles of warfare ethics fail.[1800]

V. Explainable AI (XAI) as a Technical and Legal Bridge

5.1 Concept and Key Methods

Explainable AI (XAI) offers a range of methods to keep performance metrics intact while providing explanations of the often-unexplainable decisions made by deep learning.[1801] Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) generate a particular result and explain it by building a post-hoc approximation to determine which of the input features were the most responsible for the result.[1802] Counterfactual explanations further show what minimal change in inputs would have altered the decision.[1803] In relation to the example from the ShieldAI hypothetical, such tools would trace the audit in an example such as, "Hospital grid targeted because packet signature matched 97.4% of prior hostile patterns and energy spike exceeded threshold X," *This is an example of an "explanation"* that changes a black box output into an explainable and reviewable output.

5.2 Legal Utility

XAI helps both attribution and IHL compliance. The generated explanation, under the ARSIWA framework and the Tallinn Manual 2.0, either justifies or counters the necessary link of control by showing whether the system operated within the parameters of a state's programming.[1804] Regarding reviews of Article 36 weapons, XAI logs may be submitted by the states to show the system's capability to differentiate between and make proportionality assessments to civilians and combatants. [1805] In the ShieldAI scenario, an XAI report would allow State A to prove due diligence or enable State B to demonstrate unlawful deviation—restoring the evidentiary foundation that black-box systems otherwise destroy. XAI thereby operationalises the emerging norm of "meaningful human control" by providing operators with actionable insight before and after deployment.[1806]

5.3 Limitations

Despite its merits, XAI has its limitations. Cynthia Rudin has effectively convinced us that post-hoc explanations can oversimplify or misrepresent a model's true behaviour, leaving us with an illusion of transparency.[1807] In critical military scenarios, XAI outputs might be viewed by states as sensitive, and national security concerns may preclude their being made public. Explanation-based methods could be used by adversaries to reverse engineer, and ultimately, defeat the system.[1808] Thus, while XAI narrows the accountability gap, it cannot eliminate the inherent tension between algorithmic complexity and legal certainty.

VI. Reform Proposals and the Way Forward

6.1 New Attribution Standard

International law needs to go beyond the control tests of Nicaragua and Tadić. A standard of "*constructive control*" should be adopted: any state that militarizes or neglects to control an autonomous AI system that can inflict wrongful acts internationally, is liable for the consequences of the system, irrespective of human control.[1809] Such an approach is consistent with the risk-based due-diligence

[1800] Rebecca Crootof, War, Responsibility, and Killer Robots, 40 N.C. J. Int'l L. 909, 932–35 (2015).

[1801] Finale Doshi-Velez & Been Kim, Towards a Rigorous Science of Interpretable Machine Learning, arXiv:1702.08608, at 2–4 (2017).

[1802] Marco Tulio Ribeiro et al., "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1135 (2016) (LIME); Scott M. Lundberg & Su-In Lee, A Unified Approach to Interpreting Model Predictions, 30 Advances Neural Info. Processing Sys. 4765 (2017) (SHAP).

[1803] Sandra Wachter et al., Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR, 31 Harv. J.L. & Tech. 841, 852–55 (2018).

[1804] Michael N. Schmitt, Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations 30–32 (2d ed. 2017).

[1805] Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), art. 36, June 8, 1977, 1125 U.N.T.S. 3.

[1806] Rebecca Crootof, War, Responsibility, and Killer Robots, 40 N.C. J. Int'l L. 909, 932–35 (2015)

[1807] Cynthia Rudin, Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, 1 Nature Mach. Intell. 206, 208–10 (2019).

[1808] See generally Heather M. Roff, The Strategic Implications of Autonomous Weapons, 25 Ethics & Int'l Aff. 1, 12–15 (2019).

[1809] See Rebecca Crootof, War, Responsibility, and Killer Robots, 40 N.C. J. Int'l L. 909, 932–35 (2015) (proposing risk-based models for autonomous systems).

INDIAN JOURNAL OF LEGAL REVIEW [IJLR – IF SCORE – 7.58]
VOLUME 6 AND ISSUE 2 OF 2026
APIS – 3920 - 0001 *(and)* ISSN - 2583-2344
Published by
Institute of Legal Education
https://iledu.in

obligations that exist in some form in the fields of environmental and cyber law. In the ShieldAI example, State A would be liable for the attack on the hospital-grid solely for having deployed the system.

6.2 Legal Duty of Explainability

States must accept a binding duty of explainability. XAI techniques (LIME, SHAP, counterfactual logs) would be mandated at the design, training, and deployment stages of any military AI system.[1810] Not having such audit trails and the continued failure to disclose them would allow for a rebuttable presumption of liability under ARSIWA. This duty may become a part of customary international law due to state practice and *opinio juris*, or it may be included in a new law in relation to the Convention on Certain Conventional Weapons.

6.3 Shifting the Burden of Proof and Explainability-by-Design

Once harm is evidenced, the responsibility shifts to the deploying state to demonstrate—through XAI documents—compliance with legal operational boundaries and comprehensive adherence to all possible precautionary measures. [1811] Procurement contracts and national regulations must embed Explainability-by-Design requirements, ensuring transparency is non-negotiable rather than optional.

6.4 Institutional Solution

A small but useful measure is the establishment of an AI Accountability Working Group under the CCW framework or as a standing UN GGE subsidiary body. This group would draft technical norms for XAI certification, handle incident reporting confidentiality, and mediate attribution disputes.[1812] Such an institution would bridge the current normative vacuum without requiring a full treaty.

These types of reform are also pragmatic, technically possible, and normatively justified. They maintain state sovereignty, and, even when the weapons are not, provide for accountability to remain human-centred.

VII. Conclusion

In the current international legal system, assigning accountability to opaque and autonomous AI systems for criminal actions is impossible. The black box problem, which is one of the most significant issues in understanding autonomous systems, acts to make the necessary legal elements of intent, control, and causation, which are core to ARSIWA, the Nicaragua and Tadić tests, the Tallinn Manual 2.0, and the Additional Protocol I, Article 36, unworkable when it is the algorithm that is deciding the actions.[1813] In the ShieldAI hypothetical, both the deploying state and the victim state are blocked from meeting the evidentiary or doctrinal thresholds of traditional attribution, which creates a gaping structural accountability void that undermines the entire construct of state accountability and international humanitarian law.

This paper has demonstrated that while Explainable AI techniques such as LIME, SHAP, and counterfactual logs offer valuable audit trails and partial mitigation, they cannot eliminate the fundamental tension between algorithmic complexity and legal certainty.[1814] Technological responses fall short, and the doctrinal analysis in Parts II–IV combined with the reform proposals in Part VI, demonstrate

[1810] Sandra Wachter et al., Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR, 31 Harv. J.L. & Tech. 841, 852–55 (2018).

[1811] Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), art. 36, June 8, 1977, 1125 U.N.T.S. 3.

[1812] See generally U.N. Group of Governmental Experts on Lethal Autonomous Weapons Systems, Report of the 2024 Meeting, U.N. Doc. CCW/GGE.1/2024/3 (2024) (recommending ongoing institutional mechanisms).

[1813] Articles on Responsibility of States for Internationally Wrongful Acts, G.A. Res. 56/83, Annex, arts. 4–8, U.N. Doc. A/RES/56/83 (Dec. 12, 2001); Military and Paramilitary Activities in and Against Nicaragua (Nicar. v. U.S.), Judgment, 1986 I.C.J. 14, ¶¶ 109–15 (June 27); Prosecutor v. Tadić, Case No. IT-94-1-A, Judgment, ¶¶ 98–145 (Int'l Crim. Trib. for the Former Yugoslavia July 15, 1999); Michael N. Schmitt, Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations 29–35, 84–88 (2d ed. 2017); Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), art. 36, June 8, 1977, 1125 U.N.T.S. 3.

[1814] Cynthia Rudin, Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, 1 Nature Mach. Intell. 206, 208–10 (2019); Rebecca Crootof, War, Responsibility, and Killer Robots, 40 N.C. J. Int'l L. 909, 932–35 (2015).

that international law must move past its anthropocentric first principles. "*Constructive control*", a binding legal duty to explain, a shift in burden, and the establishment of an AI Accountability Working Group provide the much needed clarity to make progress.

The theoretical contribution of this study lies in bridging the disciplinary divide: it translates technical opacity into precise legal failure points and converts XAI scholarship into actionable normative proposals. As warfare becomes increasingly algorithmic, the rule of law faces a stark choice—adapt or become irrelevant. Accountability must remain human-centred, even when the weapons are not. Only by embedding transparency and explainability into both code and custom can international law preserve its moral and practical authority in the age of autonomous cyber conflict.