Autonomous Cyber Weapons and International Humanitarian Law: New Protective Architectures for Human Security

Inaugural-Dissertation
zur
Erlangung der Doktorwürde
einer Hohen Rechtswissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von

Hassan Khoder, LL.M.

Referent:	**Professor Dr. Dr. h. c. Dr. h. c. Stephan Hobe, L.L.M.**
Korreferent:	**Professor Dr. Burkhard Schöbener**

Preface

This dissertation makes one argument. International humanitarian law cannot govern autonomous cyber weapons — not because IHL's rules are wrong, but because IHL's compliance architecture presupposes something that autonomous cyber weapons eliminate by design: a human being present at the moment of lethal targeting. A human who can distinguish civilians from combatants. A human who can weigh proportionality. A human who can cancel an attack.

Stuxnet, Industroyer, and the Ukraine electricity attacks — the three operations documented in this dissertation — each removed that human from the causal chain by design, not accidentally. That is a structural incompatibility. It cannot be resolved through better State practice or stricter weapons review. The factual precondition for IHL's operation at the moment of targeting is absent. This dissertation's response is *"Human Security by Design"*: relocate IHL's precautionary obligations to the design phase of the weapon, where a human decision-maker still exists.

The International Autonomous Weapons Control Organization proposed in Chapter 6 is the institutional mechanism that makes it real. The technical and doctrinal propositions through which this argument is established are set out in the Introduction.

The question arose not in the abstract but from sustained engagement with the documented operational histories of Stuxnet, Industroyer, and the Ukraine electricity attacks, systems whose designers, deployers, and commanders had in each case authorized a weapon whose subsequent behavior none of them controlled, predicted, or, in some instances, understood. The international humanitarian law literature I consulted in analyzing these cases was sophisticated and extensive. It was also, I came to believe, insufficient, not because its principles were wrong, but

because its foundational temporal assumptions were being structurally defeated by weapons that disaggregate the moment of human decision from the moment of lethal effect by days, months, or years.

The existing scholarship had diagnosed that difficulty with considerable precision. What it had not yet done was name the structural reason for it, and draw the institutional consequence that reason demands. This dissertation attempts to do both.

This dissertation was written at the Faculty of Law of the University of Cologne under the supervision of Professor Dr. Dr. h.c. Dr. h.c. Stephan Hobe. I am grateful to Professor Hobe for accepting this project, for his patient engagement with its arguments, and for his directness and intellectual honesty. The guidance he provided shaped this work in ways that are visible on every page.

I am equally grateful to Professor Dr. Burkhard Schöbener for agreeing to serve as second examiner (*Korreferent*) for this dissertation. His willingness to bring his expertise in public international law to bear on work that sits at the intersection of cyber operations, autonomous systems, and the laws of armed conflict is a generosity I did not take for granted. Any errors of law, fact, or judgment that remain are mine alone.

Hassan Khoder
July 2026

Table of Abbreviations

Abbreviation	Full Term
AI	Artificial Intelligence
AP I	Additional Protocol I to the Geneva Conventions of 1949
AP II	Additional Protocol II to the Geneva Conventions of 1949
ACW	Autonomous Cyber Weapon(s)
ATO	Algorithmic Transparency Obligation
IAWCO	International Autonomous Weapons Control Organization
ARSIWA	Articles on the Responsibility of States for Internationally Wrongful Acts (2001)
AWS	Autonomous Weapon System(s)
BWC	Biological Weapons Convention
CCDCOE	Cooperative Cyber Defence Centre of Excellence (NATO)
CCW	Convention on Certain Conventional Weapons
CIL	Customary International Law
CTBT	Comprehensive Nuclear-Test-Ban Treaty

CWC	Chemical Weapons Convention
ECtHR	European Court of Human Rights
EU	European Union
GA / G.A.	General Assembly (United Nations)
GC I	Geneva Convention I (for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field)
GC II	Geneva Convention II (for the Amelioration of the Condition of Wounded, Sick and Ship-wrecked Members of Armed Forces at Sea)
GC III	Geneva Convention III (relative to the Treat-ment of Prisoners of War)
GC IV	Geneva Convention IV (relative to the Protec-tion of Civilian Persons in Time of War)
GGE	Group of Governmental Experts
IAC	International Armed Conflict
IACtHR	Inter-American Court of Human Rights
IAEA	International Atomic Energy Agency
ICC	International Criminal Court
ICCPR	International Covenant on Civil and Political Rights

ICJ	International Court of Justice
ICRC	International Committee of the Red Cross
ICTR	International Criminal Tribunal for Rwanda
ICTY	International Criminal Tribunal for the former Yugoslavia
IHL	International Humanitarian Law
ILC	International Law Commission
LAWS	Lethal Autonomous Weapons Systems
NATO	North Atlantic Treaty Organization
NIAC	Non-International Armed Conflict
NPT	Treaty on the Non-Proliferation of Nuclear Weapons
NSAG	Non-State Armed Group
OEWG	Open-Ended Working Group
OPCW	Organisation for the Prohibition of Chemical Weapons
S.C.	Security Council (United Nations)
SIPRI	Stockholm International Peace Research Institute
TPNW	Treaty on the Prohibition of Nuclear Weapons

UNDRIP	United Nations Declaration on the Rights of Indigenous Peoples
UNGPs	UN Guiding Principles on Business and Human Rights
UNIDIR	United Nations Institute for Disarmament Research
VCLT	Vienna Convention on the Law of Treaties

This table lists abbreviations that recur throughout the dissertation. Case-specific short forms introduced for a single authority (e.g., a case nickname defined once via "[hereinafter ...]" and used only within the surrounding footnotes) are not repeated here; they are defined at first use in the footnotes themselves, consistent with standard practice.

Table of content

Bibliography ..... **xv**

Introduction **1**

I. The Operational Record: Autonomous Cyber Execution and the Absence of Human Control 2

II. Three Structural Deficiencies of the Existing Normative Framework 5

III. Methodology, Sources, and Epistemic Framework 6

IV. Structure of the Argument 7

Chapter 1 Autonomous Cyber Weapons: Technological Architecture and Legal Classification **10**

I. Definitional Framework and Technological Parameters 11

1. The Automation–Autonomy Continuum and IHL 11

2. Four Technological Characteristics Distinguishing Autonomous Cyber Weapons 13

a) Propagation Dynamics 13

b) Temporal Persistence and Dormancy 14

c) Technical Irreversibility 16

d) Attribution Opacity and the Structural Accountability Deficit 17

3. The Verification Incapacity of Machine–Learning Architectures 19

II. Functional Classification of Autonomous Weapon Systems: Doctrinal Architecture and Its Limits 19

Chapter 2 Autonomous Cyber Weapons and the Structural Incapacity of International Law: The Jus Contra Bellum Dimension **24**

I. The Jus Contra Bellum Dimension: The Pre-Authorization Problem 26

II. The Prohibition on the Use of Force Under Article 2(4) UN Charter and Its Application to Autonomous Cyber Operations.....27
1. The Scale–and–Effects Test: Physical Effects Threshold and Divergent State Practice 28
2. Critical Infrastructure Attacks as "Force" under Article 2(4) UN Charter: The Qualification Threshold.....29
III. Anticipatory Self–Defense, the Caroline Conditions, and the Pre-Authorization Deficit under Article 51 UN Charter 31
IV. Countermeasures, Forcible Autonomous Cyber Operations, and the Structural Tension.....36
V. The Armed Conflict Threshold, Temporal Disaggregation, and the Article 36 AP I Review Obligation.....39
1. The Armed Conflict Threshold and Technological Neutrality 40
2. Temporal Disaggregation and the Fragmentation of Article 57's Decision Architecture.....43
3. Temporal Misalignment and the Article 36 AP I Review Obligation 45

Chapter 3 Autonomous Cyber Weapons and the Structural Incapacity of International Law: The Jus In Bello Dimension 49

I. The Violence Threshold under Article 49(1) AP I: Functional Interpretation, State Practice, and the Stuxnet Precedent.....50
1. The Violence Threshold: Functional Interpretation under Article 31(1) VCLT 51
2. State Practice Convergence on a Bounded Functional Harm Test 55
3. The Stuxnet Precedent: Precision Design and Indiscriminate Propagation 57
4. Functional Harm Threshold: State Practice Survey (2019–2025).....63

II. Distinction, Proportionality, and Precaution: Structural Incompatibility with Core IHL Targeting Obligations 69
1. Distinction: The Classification Incapacity of Algorithmic Targeting 71
2. Proportionality: Reverberating Harm and Epistemic Closure 78
3. The Cascade Problem: Network Topology and Verification Impossibility 81
4. The Interpretability Constraint: Why Technical Advances Cannot Satisfy the ATO 86
5. Military Necessity as a Putative Exception: Why It Does Not Resolve the Structural Incompatibility 89
6. Precaution under Article 57 AP I: Machine–Learning Failure Modes and Verification Impossibility 91
7. The Doctrinal Foundations of the ATO 102
8. The Algorithmic Transparency Obligation in State Practice: Three Documented Operational Breaches 107
a) The STM Kargu–2 in Libya (2020): Severance of In–Operandi Auditability and the Defeat of Article 57(2)(b) 107
b) The ZALA Lancet–3 in Ukraine (2022–present): Adversarial Distribution Shift and the Defeat of Ex-ante Interpretability under AP I 108
c) United States Project Maven: Specification Gaming and the Structural Incapacity of the Article 52(2) Military Objective Determination 110
d) Doctrinal Synthesis: Cumulative Evidentiary Grounding of the ATO and the Threshold for an Article 36 Prohibitory Conclusion 111
III. Special Protections: Medical Units, Objects Indispensable to Survival, and Works Containing Dangerous Forces 112

1. Medical Units and Digitized Healthcare Systems under Article 12 AP I 112
2. Objects Indispensable to Civilian Survival under Article 54 AP I 115

Chapter 4 Accountability and the Structural Responsibility Gap 119

I. Command Responsibility and Autonomous Cyber Weapons: Structural Obstacles and the Authorization-Stage Reorientation 120
1. The Normative Framework: Article 86 AP I and Article 28 Rome Statute 122
2. The Black Box Problem: Opacity as a Structural Legal Obstacle 124
3. Command Responsibility Reoriented: The Authorization Stage 127
4. Individual Criminal Responsibility under Article 25(3)(b) Rome Statute: Ordering the Deployment of a Tier 3 ACW System 135
II. The Accountability Gap: Structural Character and Prospective Remedies 138
1. The ICCPR Right to Life as a Cumulative Accountability Obligation 138
2. Structural Determinants of the Accountability Deficit 140
3. Structural Remedies: Treaty Revision, Institutional Design, and Normative Attribution 142
4. Synthesis: The Ex–Ante Governance Imperative 145

Chapter 5 The Human Security by Design Approach 148

I. Human Security as a Normative Framework: Definitional Contours and Digital-Age Application 149
II. Technical Safeguards and Resilient Architecture: Embedded Constraints on Autonomous Cyber Weapons 159

1. The Six Embedded Safeguards: Technical Architecture and IHL Grounding 160
2. Resilient Infrastructure Design: Legal Obligations and Technical Standards 173
III. Procedural Guarantees, Meaningful Human Control, and the Authorization Framework 176
1. Pre-Deployment Certification (Epistemic Condition): The Article 36 Review Obligation 177
2. The Attack Threshold, the Martens Clause, and the Cognitive-Agency Standard for Meaningful Human Control ... 186
a) The Martens Clause as Residual Normative Baseline ... 186
b) The Cognitive-Agency Standard for Meaningful Human Control 189
3. Meaningful Human Control: Tiered Authorization, State Practice, and the Normative Basis of Tier 3 Prohibition 191
4. Continuing Supervision, Mandatory Documentation, and the Command Responsibility Framework 192
5. Mandatory Embedded-Restraint Architecture (Cancellation Condition): Kill-Switch Implementation and Legal Exercise ... 195
6. The "Necessary and Reasonable Measures" Standard and Procurement Accountability: The Bemba Gombo Framework . 197
IV. Normative Pluralism: A Methodological Note 198
V. Implementation Architecture and Compliance Mechanism .. 199

Chapter 6 Governance Architecture and Institutional Frameworks for Autonomous Cyber Weapons 204

I. Existing Governance Structures: Cyber Operations, Autonomous Weapons, and Institutional Deficits 206
1. The UN Group of Governmental Experts: Mandate, Achievements, and Limitations 206
2. The CCW Process on Lethal Autonomous Weapons: Normative Stalemate and Prospective Reform 207

3. The Tallinn Manual Process, Regional Initiatives, and the Limits of Soft–Law Codification 208
II. Proposals for New International Legal Instruments: Three Regulatory Options 209
1. Option One: An AP to the Geneva Conventions 210
2. Option Two: A Standalone Treaty on Autonomous Cyber Weapons 214
3. The Recommended Architecture: A Sequenced Hybrid of All Three Options 216
III. Non-State Actors in Autonomous Weapons Governance: Roles, Responsibility, and Accountability 219
1. Technology Companies: Corporate Due Diligence and Public–Private Governance 220
2. Civil Society: Monitoring, Advocacy, and Prohibition Norm Entrepreneurship 224
3. The Academic and Research Community: Norm Entrepreneurship and Epistemic Authority 226
IV. Verification and Enforcement: Technical Challenges and Adaptive Mechanisms 227
1. Technical Verification Challenges: Opacity, Attribution, and Adapted Mechanisms 227
2. Enforcement Mechanisms: Graduated Sanctions, Prosecution, and Compliance Incentives 231
3. The Veto Problem and Its Three Legal Responses 234
Conclusion 237
ANNEX I Draft Convention on the Control of Autonomous Cyber Weapons 250

Bibliography

Afina, Yana, *Regional Perspectives on the Application of International Humanitarian Law to Lethal Autonomous Weapon Systems* (UNIDIR 2025), https://unidir.org.

African Charter on Human and Peoples' Rights, June 27, 1981, 21 I.L.M. 58.

Al-Dawoody, Ahmed, *Islamic Law and International Humanitarian Law: An Introduction to the Main Principles*, 99 Int'l Rev. Red Cross 995 (2017).

Albert, Réka, Hawoong Jeong & Albert-László Barabási, *Error and Attack Tolerance of Complex Networks*, 406 Nature 378 (2000).

Albright, David & Paul Brannan, *Did Stuxnet Take Out 1,000 Centrifuges at the Natanz Enrichment Plant?* (Inst. for Sci. & Int'l Sec., Dec. 22, 2010), https://isis-online.org/isis-reports/did-stuxnet-take-out-1000-centrifuges-at-the-natanz-enrichment-plant/.

Aldrich, G.H., *Customary International Humanitarian Law: An Interpretation on Behalf of the International Committee of the Red Cross*, 76 Brit. Y.B. Int'l L. 517 (2005).

Ambos, Kai. *Treatise on International Criminal Law: Foundations and General Part*. 2d ed. Oxford University Press, 2021

Amodei, Dario et al., Concrete Problems in AI Safety, arXiv:1606.06565 (June 21, 2016), https://arxiv.org/abs/1606.06565.

Anand, Alisha & Harry Deng, Towards Responsible AI in Defense: A Mapping and Comparative Analysis of AI Principles Adopted by States (UNIDIR 2023), https://unidir.org/files/2023-02/Brief-ResponsibleAI-Final.pdf.

Anand, Alisha & Ioana Puscas, *Proposals Related to Emerging Technologies in the Area of Lethal Autonomous Weapons Systems: A Resource Paper* (UNIDIR 2023), https://unidir.org/files/2023-05/UNIDIR-Proposals_Emerging_Technologies_Lethal_Autonomous_Weapons_Systems_2023.pdf.

Anderson, Kenneth & 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. Essays on Governance Studies 2013).

Anti-Personnel Mine Ban Convention, Membership (161 States Parties as of Mar. 2026), https://www.apminebanconvention.org/en/membership.

Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia and Herzegovina v. Serbia and Montenegro), Judgment, 2007 I.C.J. 43 (Feb. 26, 2007), https://www.icj-cij.org/case/91.

Armed Activities on the Territory of the Congo (Democratic Republic of the Congo v. Uganda), Merits, Judgment, 2005 I.C.J. Rep. 168 (Dec. 19, 2005).

ASEAN, *ASEAN Cybersecurity Cooperation Strategy 2021–2025* (Draft), https://dig.watch/resource/asean-cybersecurity-cooperation-strategy-draft-2021-2025 (last visited Apr. 16, 2026).

Ashford, Warwick, *NotPetya Attack Cost up to $300m, Says Maersk*, Computer Weekly, Aug. 15, 2017.

Assembly of States Parties to the Rome Statute of the International Criminal Court, *Elements of Crimes*, ICC-ASP/1/3 (pt II-B) (Sept. 9, 2002, as amended by the Review Conference of the Rome Statute, June 11, 2010), https://www.icc-cpi.int/sites/default/files/Publications/Elements-of-Crimes.pdf.

Autonomous Weapon Systems, in *The Palgrave Encyclopedia of Global Security Studies* 108 (Scott N. Romaniuk & Péter Marton eds., 2020).

Backstrom, Alan & Ian Henderson, *New Capabilities in Warfare: An Overview of Contemporary Technological Developments and the Associated Legal and Engineering Issues in Article 36 Weapons Reviews*, 94 Int'l Rev. Red Cross 483 (2012), https://doi.org/10.1017/S1816383112000707.

Bai, Yuntao et al., *Constitutional AI: Harmlessness from AI Feedback* (Anthropic, arXiv:2212.08073v1, Dec. 15, 2022), https://arxiv.org/abs/2212.08073.

Barabási, Albert-László & Réka Albert, *Emergence of Scaling in Random Networks*, 286 Science 509 (1999).

Bellinger, John B., III & William J. Haynes II, *A US Government Response to the International Committee of the Red Cross Study on Customary International Humanitarian Law*, 89 Int'l Rev. Red Cross 455 (2007).

Bellinger, John B., III & William J. Haynes II, *A US Government Response to the International Committee of the Red Cross's Customary International Humanitarian Law Study*, 46 Int'l Legal Materials 514 (2007).

Berman, P., *The ICRC's Advisory Service on International Humanitarian Law: The Challenge of National Implementation*, 36 Int'l Rev. Red Cross 338 (1996).

Bethlehem, Daniel, *Self-Defense Against an Imminent or Actual Armed Attack by Non-State Actors*, 106 Am. J. Int'l L. 769 (2012).

Biological Weapons Convention (Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on Their Destruction), Apr. 10, 1972, 1015 U.N.T.S. 163 (entered into force Mar. 26, 1975), https://ihl-databases.icrc.org/en/ihl-treaties/bwc-1972.

Blinding Laser Protocol (Protocol IV to the Convention on Certain Conventional Weapons), Oct. 13, 1995, 35 I.L.M. 1218 (entered into force July 30, 1998), https://ihl-databases.icrc.org/en/ihl-treaties/blinding-laser-protocol-1995.

Bo, Marta & Vincent Boulanin, *Retaining Human Responsibility in the Development and Use of Autonomous Weapons Systems: On Accountability for Violations of International Humanitarian Law Involving Autonomous Weapons Systems* (Stockholm Int'l Peace Rsch. Inst. 2023), https://www.sipri.org/sites/default/files/2022-10/2210_aws_human_responsibility.pdf.

Boothby, W.H., The Law of Targeting (Oxford Univ. Press 2012).

Boothby, William H., Weapons and the Law of Armed Conflict (2d ed., Oxford Univ. Press 2016).

Boothby, William H., Some Legal Challenges Posed by Remote Attack, 94 Int'l Rev. Red Cross 579 (2012).

Bothe, M., et al., International Law Protecting the Environment During Armed Conflict: Gaps and Opportunities, 92 Int'l Rev. Red Cross 569 (2010).

Bothe, M., K.J. Partsch & W.A. Solf, New Rules for Victims of Armed Conflicts: Commentary on the Two 1977 Protocols Additional to the Geneva Conventions of 1949 (2d ed., Martinus Nijhoff 2013).

Boulanin, Vincent & Marta Bo, Three Lessons on the Regulation of Autonomous Weapons Systems to Ensure Accountability for Violations of IHL, ICRC Humanitarian L. & Pol'y Blog (Mar. 2, 2023), https://blogs.icrc.org/law-and-policy/2023/03/02/three-lessons-autonomous-weapons-systems-ihl/.

Boulanin, Vincent et al., Artificial Intelligence, Strategic Stability and Nuclear Risk: Euro-Atlantic Perspectives (Stockholm Int'l Peace Research Inst. 2020).

Boulanin, Vincent, Laura Bruun & Netta Goussac, Autonomous Weapon Systems and International Humanitarian Law: Identifying Limits and the Required Type and Degree of Human-Machine Interaction, SIPRI Report (June 2021).

Boulanin, Vincent, Neil Davison, Netta Goussac & Marta Bo, Autonomous Weapon Systems and International Humanitarian Law: Findings, Recommendations and Remaining Challenges (SIPRI 2021), https://www.sipri.org/publications/2021/other-publications/autonomous-weapon-systems-and-international-humanitarian-law.

Boulanin, Vincent, Neil Davison, Netta Goussac & Moa Peldán Carlsson, Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control, SIPRI Report (June 2020).

Boylan, P.J., Review of the Convention for the Protection of Cultural Property in the Event of Armed Conflict (The Hague Convention of 1954), UNESCO Doc. CLT-93/WS/12 (1993).

Bradshaw and Others v. United Kingdom, App. No. 15653/22, Judgment (ECtHR 4th Section, July 22, 2025, final Dec. 8, 2025), available at HUDOC, https://hudoc.echr.coe.int.

Bromley, Mark, Export Controls, Human Security and Cyber-Surveillance Technology: Examining the Proposed Changes to the EU Dual-Use Regulation (SIPRI, Dec. 2017), https://www.sipri.org/sites/default/files/2018-01/sipri1712_bromley.pdf.

Bruun, Laura, Marta Bo & Netta Goussac, Compliance with International Humanitarian Law in the Development and Use of Autonomous Weapon Systems: What Does IHL Permit, Prohibit and Require? (SIPRI, Mar. 2023), https://www.sipri.org/publications/2023/policy-reports/compliance-international-humanitarian-law-development-and-use-autonomous-weapon-systems-what-does.

Bruun, Laura, Towards a Two-Tiered Approach to Regulation of Autonomous Weapon Systems: Identifying Pathways and Possible Elements (SIPRI, Aug. 2024), https://www.sipri.org/sites/default/files/2024-08/laws.pdf.

Çakmak, Cenap & Gökhan Güneysu, Exploring Foundational Convergence Between the Islamic Law of Armed Conflict and Modern International Humanitarian Law, 102 Int'l Rev. Red Cross 915 (2020).

Calcara, Antonio, NATO's Digital Modernisation: The Case of Cloud Computing (Hague Ctr. for Strategic Studies, May 2025), https://hcss.nl/wp-content/uploads/2025/05/NATOs-Digital-Modernisation-HCSS-2025-1.pdf-1.pdf.

Campaign to Stop Killer Robots, Annual Report 2023 (Apr. 2023–Mar. 2024), https://www.stopkillerrobots.org/wp-content/uploads/2024/08/SKR-Annual-Report-2023_WEB_Final.pdf.

Campbell McLachlan, The Principle of Systemic Integration and Article 31(3)(c) of the Vienna Convention, 54 Int'l & Comp. L.Q. 279 (2005).

Canadian Centre for Cyber Security, Cyber Threat Bulletin: Iranian Cyber Threat Response to U.S. Strikes (Feb. 28, 2026).

Capone, Francesca, Cyber Operations Against Critical Infrastructure and the Prohibition on Attacking Objects Indispensable to the Survival of the Civilian Population, 102 Int'l Rev. Red Cross 731 (2020).

Carlini, Nicholas & David Wagner, Towards Evaluating the Robustness of Neural Networks, 2017 IEEE Symposium on Security and Privacy 39, https://doi.org/10.1109/SP.2017.49.

Carson, T., Advancing the Legal Protection of the Environment in Relation to Armed Conflict: Protocol I's Threshold of Impermissible Environmental Damage and Alternatives, 82 Nordic J. Int'l L. 83 (2013).

Casey-Maslen, Stuart, Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on Their Destruction, U.N. Audiovisual Library Int'l L. (United Nations 2010), https://legal.un.org/avl/ha/cpusptam/cpusptam.html.

Cassese, Antonio, The Martens Clause: Half a Loaf or Simply Pie in the Sky?, 11 Eur. J. Int'l L. 187 (2000), https://www.ejil.org/pdfs/11/1/511.pdf.

CCW GGE, Chair's Summary - First 2025 Session of the GGE on LAWS, U.N. Doc. CCW/GGE.1/2025/WP.1 (Apr. 7, 2025).

CCW GGE, Chairperson's Summary, U.N. Doc. CCW/GGE.1/2020/WP.7 (Apr. 19, 2021), https://undocs.org/CCW/GGE.1/2020/WP.7.

CCW GGE, Report of the 2021 Session, U.N. Doc. CCW/GGE.1/2021/3 (2021), https://docs.un.org/en/CCW/GGE.1/2021/3.

CCW GGE, Revised Rolling Text as of 12 May 2025, https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2025)/CCW_GGE_LAWS_-Revised_rolling_text_as_of_12_May_2025.pdf.

CCW GGE, Revised Rolling Text as of 8 November 2024, https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2024)/Revised_rolling_text_as_of_8_November_2024_final.pdf.

CCW Meeting of the High Contracting Parties, Report of the 2019 Meeting, Annex III, U.N. Doc. CCW/MSP/2019/9 (Dec. 13, 2019), https://undocs.org/CCW/MSP/2019/9.

CCW, Report of the 2014 Informal Meeting of Experts on Lethal Autonomous Weapons Systems, U.N. Doc. CCW/MSP/2014/3 (June 11, 2014).

CCW, Report of the Fifth Review Conference of the High Contracting Parties to the Convention on Certain Conventional Weapons, U.N. Doc. CCW/CONF.V/10 (Dec. 23, 2016).

CDDH, Official Records of the Diplomatic Conference on the Reaffirmation and Development of International Humanitarian Law Applicable in Armed Conflicts, Geneva (1974–1977), vol. XV, CDDH/215/Rev.1 (Federal Political Department, Bern 1978).

Chairman's Statement of the 32nd ASEAN Summit (Apr. 28, 2018), https://cil.nus.edu.sg/wp-content/uploads/2019/02/2018-Chairmans-Statement-of-32nd-ASEAN-Summit.pdf.

Charter of the United Nations, June 26, 1945, 59 Stat. 1031, T.S. No. 993, https://www.un.org/en/about-us/un-charter/full-text.

Chen, Hui et al., Countermeasures in Cyberspace (GCSP, Nov. 2024), https://www.gcsp.ch/publications/countermeasures-cyberspace.

Cherepanov, Anton, Win32/Industroyer: A New Threat for Industrial Control Systems (ESET, June 12, 2017), https://web-assets.esetstatic.com/wls/2017/06/Win32_Industroyer.pdf.

China State Council, Notice of the State Council Issuing the New Generation of Artificial Intelligence Development Plan, State Council Doc. No. 35 (July 8, 2017), https://flia.org/wp-content/uploads/2017/07/A-New-Generation-of-Artificial-Intelligence-Development-Plan-1.pdf.

CISA, FBI & HHS, Joint Cybersecurity Advisory AA20-302A: Ransomware Activity Targeting the Healthcare and Public Health Sector (updated Oct. 29, 2020), https://www.cisa.gov/news-events/cybersecurity-advisories/aa20-302a.

CISA, FBI & MS-ISAC, Alert AA21-042A: Compromise of U.S. Water Treatment Facility (Feb. 12, 2021), https://www.cisa.gov/news-events/cybersecurity-advisories/aa21-042a.

CISA, FBI & U.S. Dep't of Homeland Security, Joint Cybersecurity Advisory AA21-131A, DarkSide Ransomware: Best Practices for Preventing Business Disruption from Ransomware Attacks (May 11, 2021, last rev. July 8, 2021).

CISA, FBI, NSA et al., Joint Cybersecurity Advisory AA23-335A: IRGC-Affiliated Cyber Actors Exploit PLCs in Multiple Sectors (Dec. 1, 2023, updated Dec. 18, 2024), https://media.defense.gov/2023/Dec/04/2003350920/-1/-1/0/CSA-IRGC-AFFILIATED-CYBER-ACTORS-EXPLOIT-PLCS-IN-MULTIPLE-SECTORS.PDF.

CISA, FBI, NSA, EPA, DOE & U.S. Cyber Command, Advisory AA26-097A: Iranian-Affiliated Cyber Actors Exploit Programmable Logic Controllers Across US Critical Infrastructure (Apr. 7, 2026).

CISA, NSA & FBI et al., Joint Cybersecurity Advisory AA23-144A: People's Republic of China State-Sponsored Cyber Actor Living off the Land to Evade Detection (May 24, 2023, ver. 1.1), https://www.cisa.gov/news-events/cybersecurity-advisories/aa23-144a.

CISA, NSA, FBI et al., Joint Cybersecurity Advisory AA24-038A: PRC State-Sponsored Actors Compromise and Maintain Persistent Access to U.S. Critical Infrastructure (Feb. 7, 2024), https://media.defense.gov/2024/Feb/07/2003389935/-1/-1/0/CSA-PRC-Compromise-US-Critical-Infrastructure.PDF.

CISA/NCCIC & FBI, Alert TA17-132A: Indicators Associated With WannaCry Ransomware (May 12, 2017, updated June 7, 2018), https://www.cisa.gov/news-events/alerts/2017/05/12/indicators-associated-wannacry-ransomware.

CloudSEK, A Threat Actor Landscape Assessment of ICS/OT Targeting in the 2026 Iran–US Conflict and the Scale of the Risk (Mar. 5, 2026), https://www.cloudsek.com/blog/a-threat-actor-landscape-assessment-of-ics-ot-targeting-in-the-2026-iran-us-conflict-and-the-scale-of-the-risk.

CloudSEK, AI, the Iran–US Conflict, and the Threat to US Critical Infrastructure (Mar. 6, 2026), https://www.cloudsek.com/blog/ai-the-iran-us-conflict-and-the-threat-to-us-critical-infrastructure.

Coco, Antonio, Artificial Intelligence in Armed Conflict: The Current State of International Law, Article of War (Lieber Inst. W. Point, Aug. 18, 2025), https://lieber.westpoint.edu/artificial-intelligence-armed-conflict-cycon-2025-series-introduction.

Colonial Pipeline Restarts Operations Days After Hack, Associated Press, L.A. Times, May 12, 2021.

Common Article 3 to the Geneva Conventions of 12 August 1949, 75 U.N.T.S. 31, 85, 135, 287.

Comprehensive Nuclear-Test-Ban Treaty, opened for signature Sept. 24, 1996, 35 I.L.M. 1439 (not yet in force), https://www.ctbto.org/our-mission/the-treaty.

Convention (II) with Respect to the Laws and Customs of War on Land and Its Annex: Regulations Concerning the Laws and Customs of War on Land, July 29, 1899, 32 Stat. 1803, 1 Bevans 247, https://www.legal-tools.org/doc/7879ac/pdf.

Convention (IV) Respecting the Laws and Customs of War on Land and Its Annex: Regulations Concerning the Laws and Customs of War on Land, Oct. 18, 1907, 1 Bevans 631, https://ihl-databases.icrc.org/en/ihl-treaties/hague-conv-iv-1907.

Convention for the Protection of Cultural Property in the Event of Armed Conflict, May 14, 1954, 249 U.N.T.S. 240, https://ihl-databases.icrc.org/en/ihl-treaties/hague-conv-1954.

Convention on Certain Conventional Weapons, Meeting of the High Contracting Parties, Guiding Principles Affirmed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Annex III, U.N. Doc. CCW/MSP/2019/9 (Nov. 13, 2019), https://undocs.org/CCW/MSP/2019/9.

Convention on Cluster Munitions, May 30, 2008, 2688 U.N.T.S. 39.

Convention on International Liability for Damage Caused by Space Objects, Mar. 29, 1972, 961 U.N.T.S. 187.

Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, Oct. 10, 1980, 1342 U.N.T.S. 137.

Convention on the Prohibition of Military or Any Other Hostile Use of Environmental Modification Techniques (ENMOD), May 18, 1977, 1108 U.N.T.S. 151 (entered into force Oct. 5, 1978), https://ihl-databases.icrc.org/en/ihl-treaties/enmod-1977.

Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on Their Destruction, Jan. 13, 1993, 1974 U.N.T.S. 45 (entered into force Apr. 29, 1997), https://www.opcw.org/sites/default/files/documents/CWC/CWC_en.pdf.

Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on Their Destruction, Sept. 18, 1997, 2056 U.N.T.S. 211 (entered into force Mar. 1, 1999), https://www.apminebanconvention.org/en/convention-text.

Copeland, Damian, Book Review: Natalia Jevglevskaja, International Law and Weapons Review, 13 J. Int'l Humanitarian Legal Stud. 379 (2022), https://brill.com/view/journals/ihls/13/2/article-p379_009.xml.

Copeland, Damian, Rain Liivoja & Lauren Sanders, The Utility of Weapons Reviews in Addressing Concerns Raised by Autonomous Weapon Systems, 28 J. Conflict & Security L. 285 (2023), https://doi.org/10.1093/jcsl/krac035.

Corfu Channel (United Kingdom v. Albania), Merits, Judgment, 1949 I.C.J. Rep. 4 (Apr. 9, 1949), https://www.icj-cij.org/node/103099.

Corn, G. & J.A. Schoettler, Targeting and Civilian Risk Mitigation: The Essential Role of Precautionary Measures, 223 Mil. L. Rev. 785 (2015).

Council of the European Union, Conclusions on a Framework for a Joint EU Diplomatic Response to Malicious Cyber Activities ("Cyber Diplomacy Toolbox"), Council Doc. 9916/17 (June 19, 2017), https://data.consilium.europa.eu/doc/document/ST-9916-2017-INIT/en/pdf.

Council of the European Union, EU Sanctions Map (last updated 2026), https://www.sanctionsmap.eu.

Council of the European Union, Revised Implementing Guidelines of the Cyber Diplomacy Toolbox, Council Doc. 10289/23 (June 8, 2023), https://data.consilium.europa.eu/doc/document/ST-10289-2023-INIT/en/pdf.

Council Regulation (EU) 2021/821 of 20 May 2021, Setting Up a Union Regime for the Control of Exports, Brokering, Technical Assistance, Transit and Transfer of Dual-Use Items, 2021 O.J. (L 206) 1, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32021R0821.

Council Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), 2024 O.J. (L 2024/1689).

Crawford, Kate & Vladan Joler, Anatomy of an AI System: The Amazon Echo as an Anatomical Map of Human Labor, Data and Planetary Resources (AI Now Inst. & Share Lab, Sept. 7, 2018), https://anatomyof.ai.

Crawford, Kate, Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence (Yale Univ. Press 2021).

Crootof, Rebecca, Autonomous Weapon Systems and the Limits of Analogy, 9 Harv. Nat'l Sec. J. 51 (2018).

Crosignani, Matteo, Marco Macchiavelli & André F. Silva, The Propagation of Cyberattacks Through Firms' Supply Networks (Fed. Reserve Bank of N.Y., Staff Report No. 937, 2020), https://www.newyorkfed.org/research/staff_reports/sr937.

Crucitti, Paolo, Vito Latora & Massimo Marchiori, Efficiency of Scale-Free Networks: Error and Attack Tolerance, 320 Physica A 622 (2003).

CTBT Meeting of States Signatories, Resolution Establishing the Preparatory Commission for the Comprehensive Nuclear-Test-Ban Treaty Organization, CTBT/MSS/RES/1 (Nov. 19, 1996), https://www.ctbto.org/sites/default/files/2023-10/2022_treaty_booklet_E.pdf.

Cyberattack Cost Maersk as Much as $300 Million, L.A. Times, Aug. 17, 2017.

Davison, Neil, Autonomous Weapon Systems under International Humanitarian Law, in UNODA, Perspectives on Lethal Autonomous Weapon Systems, Occasional Paper No. 30, at 9 (Nov. 2017), https://digitallibrary.un.org/record/3929837?v=pdf#files.

De Falco, Marco, Stuxnet Facts Report: A Technical and Strategic Analysis (NATO Cooperative Cyber Defense Centre of Excellence 2012).

Declaration of the United Kingdom upon Ratification of AP I (Jan. 28, 1998), https://ihl-databases.icrc.org/en/ihl-treaties/api-1977/state-parties/gb.

Desgagné, R., The Prevention of Environmental Damage in Time of Armed Conflict: Proportionality and Precautionary Measures, 3 Y.B. Int'l Humanitarian L. 109 (2000).

Dinniss, Heather Harrison, Cyber Warfare and the Laws of War (Cambridge Univ. Press 2012).

Dinstein, Y., Distinction and Loss of Civilian Protection in International Armed Conflicts, 38 Isr. Y.B. on Hum. Rts. 1 (2008).

Dinstein, Y., Legitimate Military Objectives under the Current Jus in bello, 78 Int'l L. Stud. 139 (2002).

Dinstein, Y., The ICRC Customary International Humanitarian Law Study, 36 Isr. Y.B. on Hum. Rts. 1 (2006).

Dinstein, Y., The Conduct of Hostilities under the Law of International Armed Conflict (3d ed., Cambridge Univ. Press 2016).

Dinstein, Y., The International Law of Belligerent Occupation (2d ed., Cambridge Univ. Press 2019).

Directive 2022/2555 of the European Parliament and of the Council of 14 December 2022 on Measures for a High Common Level of Cybersecurity Across the Union (NIS 2 Directive), 2022 O.J. (L 333) 80, https://eur-lex.europa.eu/eli/dir/2022/2555/oj.

DLP Forum, Cyber Operations Falling Under "Attack" in IHL (Sept. 5, 2023), https://www.dlpforum.org/2023/09/06/cyber-operations-falling-under-attack-in-ihl/.

Docherty, Bonnie, Losing Humanity: The Case Against Killer Robots (Human Rights Watch & Int'l Human Rights Clinic, Nov. 19, 2012), https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots.

Document of the Stockholm Conference on Confidence- and Security-Building Measures and Disarmament in Europe (Sept. 19, 1986), https://www.osce.org/sites/default/files/f/documents/5/d/41238.pdf.

Dörmann, K. & J. Serralvo, Common Article 1 to the Geneva Conventions and the Obligation to Prevent International Humanitarian Law Violations, 96 Int'l Rev. Red Cross 707 (2014).

Dörmann, K., Elements of War Crimes under the Rome Statute of the International Criminal Court: Sources and Commentary (ICRC/Cambridge Univ. Press 2003).

Doswald-Beck, L. ed., San Remo Manual on International Law Applicable to Armed Conflicts at Sea, prepared by international lawyers and naval experts convened by the International Institute of Humanitarian Law (adopted June 1994, Cambridge Univ. Press 1995).

Dragos, Inc., CRASHOVERRIDE: Analyzing the Threat to Electric Grid Operations (June 12, 2017), https://nsarchive.gwu.edu/sites/default/files/documents/3869008/Dragos-CRASHOVERRIDE-Analyzing-the-Threat-to.pdf.

Droege, C. & M.L. Tougas, The Protection of the Natural Environment in Armed Conflict: Existing Rules and Need for Further Legal Protection, 82 Nordic J. Int'l L. 21 (2013).

Droege, Cordula & Marie-Louise Tougas, Artificial Intelligence and Machine Learning in Armed Conflict: A Human-Centred Approach, 102 Int'l Rev. Red Cross 465 (2020).

Droege, Cordula, Get Off My Cloud: Cyber Warfare, International Humanitarian Law, and the Protection of Civilians, 94 Int'l Rev. Red Cross 533 (2012).

Drumbl, M.A., Accountability for Property Crimes and Environmental War Crimes: Prosecution, Litigation, and Development (Int'l Ctr. for Transitional Justice 2009).

Drumbl, M.A., Waging War Against the World: The Need to Move from War Crimes to Environmental Crimes, in J.E. Austin & C.E. Bruch eds., The Environmental Consequences of War 620 (Cambridge Univ. Press 2000).

Dwyer, Andrew C., The Unknowable Conflict: Tracing AI, Recognition, and the Death of the (Human) Loop, in Fabio Cristiano et al. eds., Artificial Intelligence and International Conflict in Cyberspace 21 (Routledge 2023).

Dwyer, Andrew, Unknowable Conflict: Artificial Intelligence, Recognition, and the Future of Warfare (Routledge 2022).

E-ISAC & SANS Institute, Analysis of the Cyber Attack on the Ukrainian Power Grid (Mar. 18, 2016).

Egan, B., International Law, Legal Diplomacy, and the Counter-ISIL Campaign: Some Observations, 92 Int'l L. Stud. 235 (2016).

Eggert, Linda, Autonomous Weapons Systems and Human Rights, in David Edmonds ed., AI Morality 14 (Oxford Univ. Press 2024).

Eurojust, Memorandum on Battlefield Evidence (Sept. 2020), https://www.eurojust.europa.eu/sites/default/files/assets/2020_09_14_eurojust_memorandum_on_battlefield_evidence.pdf.

European Commission & High Representative of the Union for Foreign Affairs and Security Policy, The EU's Cybersecurity Strategy for the Digital Decade, JOIN(2020) 18 final (Dec. 16, 2020), https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=JOIN:2020:18:FIN.

European Medicines Agency, Policy 0070: Publication and Access to Clinical-Trial Data (Draft for Public Consultation, June 24, 2013), https://www.ema.europa.eu/en/documents/other/draft-policy-70-publication-and-access-clinical-trial-data_en.pdf.

Falliere, Nicolas, Liam O Murchu & Eric Chien, W32.Stuxnet Dossier (Symantec Security Response ver. 1.4, Feb. 2011), https://web.archive.org/web/20150427095636/https://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/w32_stuxnet_dossier.pdf.

Faragasso, Spencer, Russian Lancet-3 Kamikaze Drone Depends on Western Parts (Inst. for Sci. & Int'l Sec., Dec. 18, 2023), https://isis-online.org/uploads/isis-reports/documents/Russian_Lancet-3_Kamikaze_Drone_Depends_on_Western_Parts_December_18_2023_FINAL.pdf.

FBI & CISA, Joint Statement: People's Republic of China (PRC) Targeting of Commercial Telecommunications Infrastructure (Nov. 13, 2024), https://www.cisa.gov/news-events/news/joint-statement-fbi-and-cisa-peoples-republic-china-prc-targeting-commercial-telecommunications.

FBI, Internet Crime Complaint Ctr., 2022 Internet Crime Report (2023), https://www.ic3.gov/AnnualReport/Reports/2022_ic3report.pdf.

Feilchenfeld, E.H., The International Economic Law of Belligerent Occupation (Carnegie Endowment for International Peace 1942).

Findlay, Trevor, Confidentiality and Verification: The IAEA and OPCW, Trust & Verify, No. 114, May–June 2004, https://www.vertic.org/media/assets/TV114.pdf.

Finnemore, Martha & Kathryn Sikkink, International Norm Dynamics and Political Change, 52 Int'l Org. 887 (1998).

FireEye Mandiant, Highly Evasive Attacker Leverages SolarWinds Supply Chain to Compromise Multiple Global Victims With SUNBURST Backdoor (Dec. 13, 2020), https://cloud.google.com/blog/topics/threat-intelligence/evasive-attacker-leverages-solarwinds-supply-chain-compromises-with-sunburst-backdoor (archived from original FireEye publication; Mandiant subsequently acquired by Google).

Fleck, D., Legal Protection of the Environment: The Double Challenge of Non-International Armed Conflict and Post-Conflict Peacebuilding, in C. Stahn, J. Iverson & J.S. Easterday eds., Environmental Protection and Transitions from Conflict to Peace: Clarifying Norms, Principles, and Practices 203 (Oxford Univ. Press 2017).

Foreign Office Minister Condemns Russia for NotPetya Attacks, U.K. Foreign & Commonwealth Off. (Feb. 14, 2018), https://www.gov.uk/government/news/foreign-office-minister-condemns-russia-for-notpetya-attacks.

France & Germany, Outline for a Normative and Operational Framework on Emerging Technologies in the Area of LAWS, U.N. Doc. CCW/GGE.1/2021/WP.5 (Sept. 27, 2021), https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Seventh_Group_of_Governmental_Experts(2021)/CCW-GGE.1-2021-WP.5_English.pdf.

Franck, Thomas M., On Proportionality of Countermeasures in International Law, 102 Am. J. Int'l L. 715 (2008).

Freeland, S., Addressing the Intentional Destruction of the Environment During Warfare Under the Rome Statute of the International Criminal Court (Intersentia 2015).

G.A. Res. 1653 (XVI), U.N. Doc. A/RES/1653(XVI) (Nov. 24, 1961), https://undocs.org/A/RES/1653(XVI).

G.A. Res. 2625 (XXV), Declaration on Principles of International Law Concerning Friendly Relations and Co-operation Among States, U.N. Doc. A/RES/2625(XXV) (Oct. 24, 1970), https://undocs.org/A/RES/2625(XXV).

G.A. Res. 377 (V), Uniting for Peace, U.N. Doc. A/RES/377(V) (Nov. 3, 1950), https://undocs.org/A/RES/377(V).

G.A. Res. 56/83, Responsibility of States for Internationally Wrongful Acts, U.N. Doc. A/RES/56/83 (Dec. 12, 2001), https://undocs.org/A/RES/56/83.

G.A. Res. 61/295, United Nations Declaration on the Rights of Indigenous Peoples, U.N. Doc. A/RES/61/295 (Sept. 13, 2007), https://www.un.org/development/desa/indigenouspeoples/wp-content/uploads/sites/19/2018/11/UNDRIP_E_web.pdf.

G.A. Res. 66/290, Follow-Up to Paragraph 143 on Human Security of the 2005 World Summit Outcome, U.N. Doc. A/RES/66/290 (Sept. 10, 2012), https://digitallibrary.un.org/record/737105/files/A_RES_66_290-EN.pdf.

G.A. Res. 73/27, U.N. Doc. A/RES/73/27 (Dec. 5, 2018), https://undocs.org/A/RES/73/27.

G.A. Res. 75/240, U.N. Doc. A/RES/75/240 (Dec. 31, 2020), https://undocs.org/A/RES/75/240.

G.A. Res. 76/300, U.N. Doc. A/RES/76/300 (July 28, 2022), https://docs.un.org/en/a/res/76/300.

G.A. Res. 78/241, Lethal Autonomous Weapons Systems, U.N. Doc. A/RES/78/241 (adopted Dec. 22, 2023, distributed Dec. 28, 2023), https://undocs.org/A/RES/78/241.

G.A. Res. 79/239, Developments in the Field of Information and Telecommunications in the Context of International Security, U.N. Doc. A/RES/79/239 (Dec. 24, 2024), https://undocs.org/A/RES/79/239.

G.A. Res. 79/62, Lethal Autonomous Weapons Systems, U.N. Doc. A/RES/79/62 (Dec. 2, 2024), https://undocs.org/A/RES/79/62.

G.A. Res. 80/57, Lethal Autonomous Weapons Systems, U.N. Doc. A/RES/80/57 (Dec. 1, 2025), https://undocs.org/en/A/RES/80/57.

G.A. Res. 80/58, Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security, U.N. Doc. A/RES/80/58 (Dec. 1, 2025), https://undocs.org/A/RES/80/58.

G7 Think7 Task Force, G7-led Principles for Military AI Governance: Imperatives and Pathways (2023), https://archive.think7.org/wp-content/uploads/2024/05/T7it_tf1_pb04.pdf.

Gabčíkovo-Nagymaros Project (Hungary v. Slovakia), Judgment, 1997 I.C.J. 7 (Sept. 25, 1997), https://www.icj-cij.org/case/92.

Gayanashagowa (Great Law of Peace) (traditional oral constitution of the Haudenosaunee Confederacy).

Gaynor, K.M., et al., War and Wildlife: Linking Armed Conflict to Conservation, 14 Frontiers Ecology & Env't 533 (2016).

Geiss, R., The Obligation to Respect and to Ensure Respect for the Conventions, in A. Clapham, P. Gaeta & M. Sassòli eds., The 1949 Geneva Conventions: A Commentary 111 (Oxford Univ. Press 2015).

Geneva Convention for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field (Geneva Convention I), Aug. 12, 1949, 75 U.N.T.S. 31, https://ihl-databases.icrc.org/en/ihl-treaties/gci-1949.

Geneva Convention for the Amelioration of the Condition of Wounded, Sick and Shipwrecked Members of Armed Forces at Sea (Geneva Convention II), Aug. 12, 1949, 75 U.N.T.S. 85, https://ihl-databases.icrc.org/en/ihl-treaties/gcii-1949.

Geneva Convention Relative to the Protection of Civilian Persons in Time of War (Geneva Convention IV), Aug. 12, 1949, 75 U.N.T.S. 287, https://ihl-databases.icrc.org/en/ihl-treaties/gciv-1949.

Geneva Convention Relative to the Treatment of Prisoners of War (Geneva Convention III), Aug. 12, 1949, 75 U.N.T.S. 135, https://ihl-databases.icrc.org/en/ihl-treaties/gciii-1949.

Gill, T.D. & D. Fleck, The Handbook of the International Law of Military Operations (2d ed., Oxford Univ. Press 2015).

Gillard, E.C., Proportionality in the Conduct of Hostilities: The Incidental Harm Side of the Assessment (Chatham House, Dec. 2018).

Gisel, L. ed., The Principle of Proportionality in the Rules Governing the Conduct of Hostilities under International Humanitarian Law, Report of the International Expert Meeting, Quebec, 22–23 June 2016 (ICRC 2018).

Gisel, L., Relevant Incidental Harm for the Proportionality Principle, in ICRC/College of Europe, Urban Warfare, Proceedings of the 16th Bruges Colloquium, 15–16 October 2015, Collegium No. 46, at 118 (Autumn 2016).

Gisel, L., The Relevance of Revenue-Generating Objects in Relation to the Notion of Military Objective, in ICRC/College of Europe, The Additional Protocols at 40: Achievements and Challenges, Proceedings of the 18th Bruges Colloquium, 19–20 October 2017, Collegium No. 48, at 139 (Autumn 2018).

Gisel, L., The Use of Explosive Weapons in Densely Populated Areas and the Prohibition of Indiscriminate Attacks, in E. Greppi ed., Conduct of Hostilities: The Practice, the Law and the Future 100 (Franco Angeli 2015).

Gisel, Laurent, Tilman Rodenhäuser & Knut Dörmann, Twenty Years On: International Humanitarian Law and the Protection of Civilians Against the Effects of Cyber Operations During Armed Conflicts, 102 Int'l Rev. Red Cross 287 (2020).

Goldblat, J., The Biological Weapons Convention: An Overview, 37 Int'l Rev. Red Cross 251 (1997).

Goodfellow, Ian et al., Generative Adversarial Nets, in 27 Advances in Neural Information Processing Systems 2672 (Z. Ghahramani et al. eds., 2014), https://papers.nips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

Goodfellow, Ian J., Jonathon Shlens & Christian Szegedy, Explaining and Harnessing Adversarial Examples, 3d Int'l Conf. on Learning Representations (ICLR 2015), arXiv:1412.6572v3 (Mar. 20, 2015), https://arxiv.org/abs/1412.6572.

Goodman, R., The Obama Administration and Targeting "War-Sustaining" Objects in Non-International Armed Conflict, 110 Am. J. Int'l L. 663 (2016).

Graham, Mary, Some Thoughts About the Philosophical Underpinnings of Aboriginal Worldviews, 45 Austl. Humanities Rev. (Nov. 2008), http://australianhumanitiesreview.org/2008/11/01/some-thoughts-about-the-philosophical-underpinnings-of-aboriginal-worldviews/.

Greenberg, Andy, Sandworm: A New Era of Cyberwar and the Hunt for the Kremlin's Most Dangerous Hackers (Doubleday 2019).

Greenberg, M.I., K.J. Sexton & D. Vearrier, Sea-Dumped Chemical Weapons: Environmental Risk, Occupational Hazard, 54 Clinical Toxicology 79 (2016).

Group of Governmental Experts of the High Contracting Parties to the Convention on Certain Conventional Weapons, Draft Report of the 2019 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, U.N. Doc. CCW/GGE.1/2019/CRP.1/Rev.2 (Aug. 21, 2019), https://undocs.org/CCW/GGE.1/2019/CRP.1/Rev.2.

Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security, Report of the Group of Governmental Experts, U.N. Doc. A/76/135 (July 14, 2021), https://documents.un.org/doc/undoc/gen/n21/166/74/pdf/n2116674.pdf.

Group of Governmental Experts on Lethal Autonomous Weapons Systems, Report of the 2023 Session, U.N. Doc. CCW/GGE.1/2023/2 (advance version, 2023).

Guzman, Andrew T., A Compliance-Based Theory of International Law, 90 Calif. L. Rev. 1823 (2002).

Haas, Peter M., Introduction: Epistemic Communities and International Policy Coordination, 46 Int'l Org. 1 (1992).

Haataja, Samuli, Autonomous Cyber Capabilities and Attribution, in Rain Liivoja & Anna Väljataga eds., Autonomous Cyber Capabilities Under International Law 279 (NATO CCDCOE 2021).

Hague Centre for Strategic Studies & CyberPeace Institute, Cyber Transparency Value Chain: From Awareness and Understanding to Attribution, Monitoring and Sanctioning (2024).

Hallaq, Wael B., An Introduction to Islamic Law (Cambridge Univ. Press 2009).

Hambling, David, Russian Loitering Munition Racks Up Kills But Shows Limitations, Forbes, Dec. 1, 2022, https://www.forbes.com/sites/davidhambling/2022/12/01/russian-loitering-munition-racks-up-kills-but-shows-limitations/.

Hamdan v. Rumsfeld, 548 U.S. 557 (2006).

Hanson, T., et al., Warfare in Biodiversity Hotspots, 23 Conservation Biology 578 (2009).

Hart, J., Background to Selected Environmental and Human Health Effects of Chemical Warfare Agents, in T.A. Kassim & D. Barceló eds., Environmental Consequences of War and Aftermath 2 (Springer 2009).

Hegseth, Pete, Secretary of Defense, Statement on Iranian Military Command Degradation, as reported in contemporaneous press coverage of Operation Epic Fury (Feb. 28–Mar. 5, 2026).

Heintschel von Heinegg, W. & M. Donner, New Developments in the Protection of the Natural Environment in Naval Armed Conflicts, 37 German Y.B. Int'l L. 281 (1994).

Heller, K.J. & J.C. Lawrence, The First Ecocentric Environmental War Crime: The Limits of Article 8(2)(b)(iv) of the Rome Statute, 20 Geo. Int'l Envtl. L. Rev. 61 (2007).

Hellquist, Elin, Regional Organizations and Sanctions Against Members: Explaining the Different Trajectories of the African Union, the League of Arab States, and the Association of Southeast Asian Nations, KFG Working Paper No. 59 (Freie Universität Berlin, Jan. 2014), https://userpage.fu-berlin.de/kfgeu/kfgwp/wpseries/WorkingPaperKFG_59.pdf.

Henckaerts, J.M., Customary International Humanitarian Law: A Rejoinder to Judge Aldrich, 76 Brit. Y.B. Int'l L. 525 (2005).

Henckaerts, J.M., Customary International Humanitarian Law: A Response to US Comments, 89 Int'l Rev. Red Cross 473 (2007).

Henckaerts, J.M., The ICRC Customary International Humanitarian Law Study: A Rejoinder to Professor Dinstein, 37 Isr. Y.B. on Hum. Rts. 259 (2007).

Henckaerts, J.M. & D. Constantin, Protection of the Natural Environment, in A. Clapham & P. Gaeta eds., The Oxford Handbook of International Law in Armed Conflict 469 (Oxford Univ. Press 2014).

Henckaerts, J.M. & L. Doswald-Beck eds., Customary International Humanitarian Law, Vol. I: Rules and Vol. II: Practice (ICRC/Cambridge Univ. Press 2005, reprinted 2009).

Henderson, Ian, The Contemporary Law of Targeting: Military Objectives, Proportionality and Precautions in Attack Under Additional Protocol I (Martinus Nijhoff 2009).

Hendrycks, Dan & Thomas G. Dietterich, Benchmarking Neural Network Robustness to Common Corruptions and Perturbations, in Proceedings of the International Conference on Learning Representations 1 (2019), https://openreview.net/pdf?id=HJz6tiCqYm.

Heyns, Christof, Human Rights and the Use of Autonomous Weapons Systems During Domestic Law Enforcement, 38 Hum. Rts. Q. 350 (2016).

Hirsch, Theodore, The IAEA Additional Protocol: What It Is and Why It Matters, 11 Nonproliferation Rev. 140 (2004), https://www.nonproliferation.org/wp-content/uploads/npr/113hirsch.pdf.

Hoffberger-Pippan, Elisabeth & Christian Pippan, Inherited Obstruction: The Complaint Procedure Under the Biological Weapons Convention and the UN Charter Rule on Obligatory Abstention in the Security Council, 30 J. Conflict & Security L. 121 (2025), https://academic.oup.com/jcsl/article/30/1/121/7974355.

Hoffman, U. & P. Rapillard, Do No Harm in Mine Action: Why the Environment Matters, 19 J. ERW & Mine Action 4 (2015).

Hulme, K., Natural Environment, in E. Wilmshurst & S. Breau eds., Perspectives on the ICRC Study on Customary International Humanitarian Law 204 (Cambridge Univ. Press 2007).

Hulme, K., Taking Care to Protect the Environment Against Damage: A Meaningless Obligation?, 92 Int'l Rev. Red Cross 675 (2010).

Hulme, K., War Torn Environment: Interpreting the Legal Threshold (Martinus Nijhoff 2004).

Human Rights Committee, General Comment No. 31 on the Nature of the General Legal Obligation Imposed on States Parties to the Covenant, U.N. Doc. CCPR/C/21/Rev.1/Add.13 (May 26, 2004).

Human Rights Committee, General Comment No. 36 on Article 6 of the International Covenant on Civil and Political Rights, on the Right to Life, U.N. Doc. CCPR/C/GC/36 (Sept. 3, 2019).

IAEA, Model Protocol Additional to the Agreement(s) Between State(s) and the International Atomic Energy Agency for the Application of Safeguards, IAEA Doc. INFCIRC/540(Corrected) (Sept. 1997), https://www.iaea.org/sites/default/files/infcirc540.pdf.

ICRC, Advisory Service on International Humanitarian Law: 2017 Treaty on the Prohibition of Nuclear Weapons, Factsheet (2018).

ICRC, Commentary on the First Geneva Convention (2016).

ICRC, Guidelines for Military Manuals and Instructions on the Protection of the Environment in Times of Armed Conflict (1994), annexed to U.N. GAOR, United Nations Decade of International Law: Report of the Secretary-General, U.N. Doc. A/49/323 (Aug. 19, 1994).

ICRC, Nuclear Weapons and International Humanitarian Law, Information Note No. 4 (ICRC 2013).

ICRC, Strengthening Legal Protection for Victims of Armed Conflicts, Report submitted to the 31st International Conference of the Red Cross and Red Crescent (ICRC, Oct. 2011).

ICRC, When Rain Turns to Dust: Understanding and Responding to the Combined Impact of Armed Conflicts and the Climate and Environment Crisis on People's Lives (ICRC 2020).

ILA (International Law Association) Study Group on the Conduct of Hostilities in the 21st Century, The Conduct of Hostilities and International Humanitarian Law: Challenges of 21st Century Warfare, 93 Int'l L. Stud. 322 (2017).

ILC (International Law Commission), Preliminary Report on the Protection of the Environment in Relation to Armed Conflicts, submitted by Marie G. Jacobsson, Special Rapporteur, U.N. Doc. A/CN.4/674 (May 30, 2014).

ILC, Draft Articles on the Effects of Armed Conflicts on Treaties (2011), reproduced in U.N. GAOR, Report of the International Law Commission: Sixty-Third Session (26 April–3 June and 4 July–12 August 2011), U.N. Doc. A/66/10 (2011).

ILC, Draft Conclusions on Identification of Customary International Law, U.N. Doc. A/73/10 (2018).

ILC, Draft Principles on the Protection of the Environment in Relation to Armed Conflicts (2019), reproduced in U.N. GAOR, Report of the International Law Commission: Seventy-First Session (29 April–7 June and 8 July–9 August 2019), U.N. Doc. A/74/10, at 209–296 (2019).

ILC, First Report on Protection of the Environment in Relation to Armed Conflicts, by Marja Lehto, Special Rapporteur, U.N. Doc. A/CN.4/720 (Apr. 30, 2018).

ILC, Second Report on Protection of the Environment in Relation to Armed Conflicts, by Marja Lehto, Special Rapporteur, U.N. Doc. A/CN.4/728 (Mar. 27, 2019).

ILC, Second Report on the Protection of the Environment in Relation to Armed Conflicts, submitted by Marie G. Jacobsson, Special Rapporteur, U.N. Doc. A/CN.4/685 (May 28, 2015).

ILC, Third Report on the Protection of the Environment in Relation to Armed Conflicts, submitted by Marie G. Jacobsson, Special Rapporteur, U.N. Doc. A/CN.4/700 (June 3, 2016).

Int'l Comm. Red Cross, Autonomous Weapon Systems and International Humanitarian Law (Mar. 2, 2026), https://www.icrc.org/sites/default/files/2026-03/4896_002_Autonomous_Weapons_Systems_-_IHL-ICRC.pdf.

Int'l Comm. Red Cross, International Humanitarian Law and the Challenges of Contemporary Armed Conflicts, Doc. 34IC/24/10.6 (ICRC 2024), https://rcrcconference.org/app/uploads/2024/10/34IC-IHL-Challenges-Report-EN.pdf.

Int'l Comm. Red Cross, The Principle of Distinction (Mar. 2023), https://www.icrc.org/sites/default/files/wysiwyg/war-and-law/03_distinction-0.pdf.

Int'l Conf. of the Red Cross & Red Crescent (34th), Resolution 2: Protecting Civilians and Other Protected Persons and Objects Against the Potential Human Cost of ICT Activities During Armed Conflict, Doc. 34IC/24/R2 (Oct. 2024), https://rcrcconference.org/app/uploads/2024/10/34IC_R2-ICT-EN.pdf.

Int'l L. Comm'n, Articles on Responsibility of States for Internationally Wrongful Acts, with Commentaries, in Report of the International Law Commission on the Work of Its Fifty-Third Session, U.N. GAOR, 56th Sess., Supp. No. 10, U.N. Doc. A/56/10 (2001), annexed to G.A. Res. 56/83, U.N. Doc. A/RES/56/83 (Dec. 12, 2001), https://legal.un.org/ilc/texts/instruments/english/draft_articles/9_6_2001.pdf.

Int'l L. Comm'n, Draft Articles on Prevention of Transboundary Harm from Hazardous Activities, with Commentaries, in Report of the International Law Commission on the Work of Its Fifty-Third Session, U.N. GAOR, 56th Sess., Supp. No. 10, U.N. Doc. A/56/10 (2001), https://legal.un.org/ilc/texts/instruments/english/draft_articles/9_7_2001.pdf.

Int'l L. Comm'n, Draft Articles on the Responsibility of International Organizations, U.N. Doc. A/66/10 (2011), https://legal.un.org/ilc/texts/instruments/english/draft_articles/9_11_2011.pdf.

Int'l L. Comm'n, Fragmentation of International Law: Difficulties Arising from the Diversification and Expansion of International Law, U.N. Doc. A/CN.4/L.682 (Apr. 13, 2006), https://legal.un.org/ilc/documentation/english/a_cn4_l682.pdf.

Int'l Monetary Fund, Global Financial Stability Report: The Last Mile: Financial Vulnerabilities and Risks (Apr. 2024), https://www.imf.org/en/Publications/GFSR/Issues/2024/04/16/global-financial-stability-report-april-2024.

International Covenant on Civil and Political Rights, Dec. 16, 1966, 999 U.N.T.S. 171 (entered into force Mar. 23, 1976), https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights.

Jackson, Miles. *Causation and the Legal Character of Command Responsibility After* Bemba *at the International Criminal Court*, 20 J. Int'l Crim. Just. 437 (2022).

Jain, Sarthak, and Byron C. Wallace. "Attention Is Not Explanation." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3543–48. Association for Computational Linguistics, 2019. https://aclanthology.org/N19-1357.

Jennings, R.Y., The Caroline and McLeod Cases, 32 Am. J. Int'l L. 82 (1938).

Jevglevskaja, Natalia, International Law and Weapons Review: Emerging Military Technology Under the Law of Armed Conflict (Cambridge Univ. Press 2021).

Jevglevskaja, Natalia, Weapons Review Obligation Under Customary International Law, 94 Int'l L. Stud. 186 (2018).

Johns, Fleur, Digital Humanitarianism: Interfaces, International Law and Politics (Oxford Univ. Press 2023).

Joint Chiefs of Staff, Joint Publication 3-60: Joint Targeting (Jan. 31, 2013), https://www.justsecurity.org/wp-content/uploads/2015/06/Joint_Chiefs-Joint_Targeting_20130131.pdf.

Kamali, Mohammad Hashim, Maqāṣid al-Sharīʿah: The Objectives of Islamic Law, 38 Islamic Stud. 193 (1999).

Kaminska, Monica, Fabio Cristiano & Dennis Broeders, Limiting Viral Spread: Automated Cyber Operations and the Principles of Distinction and Discrimination in the Grey Zone, in 13th International Conference on Cyber Conflict: Going Viral 60 (NATO CCDCOE 2021).

Kassim, T.A. & D. Barceló eds., Environmental Consequences of War and Aftermath (Springer-Verlag 2009).

Kastelic, Andraz, Due Diligence in Cyberspace: Normative Expectations of Reciprocal Protection of International Legal Rights (UNIDIR 2023), https://unidir.org/wp-content/uploads/2023/05/SecurityTechnology_DueDiligence_Report.pdf.

Koh, Harold Hongju, International Law in Cyberspace, 54 Harv. Int'l L.J. Online 1 (Dec. 2012), https://harvardilj.org/wp-content/uploads/2012/12/HILJ-Online_54_Koh.pdf.

Kott, Alexander et al., Toward Intelligent Autonomous Agents for Cyber Defense: Report of the 2017 Workshop by the NATO Research Group IST-152/RTG (U.S. Army Research Lab. Special Report ARL-SR-0395, Apr. 2018).

Krakovna, Victoria et al., Avoiding Side Effects in Complex Environments, 34 Advances in Neural Info. Processing Sys. (NeurIPS) 1 (2020).

Kreß, Claus, Some Reflections on the International Legal Framework Governing Transnational Armed Conflicts, 15 J. Conflict & Sec. L. 245 (2010).

Kulkarni, Ajay et al., A Review of Cybersecurity Incidents in the Food and Agriculture Sector (Mar. 12, 2024), https://arxiv.org/pdf/2403.08036.pdf.

Kwik, Jonathan, Iterative Assessment for Military Artificial Intelligence (AI) Systems (Asser Inst. Research Paper 2025-03, 2025), https://www.asser.nl/media/798036/5338856.pdf.

Kwik, Jonathan, Marten Zwanenburg, Cheya Took & Joep Aarts, Controlling Military Artificial Intelligence: Harnessing Rules of Engagement and Military Directives, Asser Policy Brief 2025-01 (Asser Inst. Feb. 2025).

La Haye, E., War Crimes in Internal Armed Conflicts (Cambridge Univ. Press 2008).

Lancet Loitering Munition System, Russia, Army Technology (Sept. 11, 2024), https://www.army-technology.com/projects/lancet-loitering-munition-system-russia/.

Langner, Ralph, To Kill a Centrifuge: A Technical Analysis of What Stuxnet's Creators Tried to Achieve (Nov. 2013), https://cyber-peace.org/wp-content/uploads/2013/06/To-kill-a-centrifuge.pdf.

Le Billon, P., Wars of Plunder: Conflicts, Profits and the Politics of Resources (Hurst 2012).

LeCun, Yann et al., Deep Learning, 521 Nature 436 (2015).

Legal Consequences of the Construction of a Wall in the Occupied Palestinian Territory, Advisory Opinion, 2004 I.C.J. Rep. 136 (July 9, 2004), https://www.icj-cij.org/case/131.

Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 1996 I.C.J. Rep. 226 (July 8, 1996), https://www.icj-cij.org/case/95.

Letter from Daniel Webster, U.S. Secretary of State, to Lord Ashburton (Apr. 24, 1841), reprinted in 2 John Bassett Moore, A Digest of International Law 412 (Gov't Printing Off. 1906).

Levie, H.S., Protection of War Victims: Protocol I to the 1949 Geneva Conventions, vol. 3 (Oceana Publications 1980).

Lewis, Dustin A. & Hannah Sweeney, Exercising Cognitive Agency: A Legal Framework Concerning Natural and Artificial Intelligence in Armed Conflict (HLS PILAC Legal Concept Paper, Jan. 2025), https://pilac.law.harvard.edu.

Lucas, George R., Law, Ethics and Emerging Military Technologies: Confronting Disruptive Innovation (Routledge 2023).

Lundberg, M.A., The Plunder of Natural Resources During War: A War Crime?, 39 Geo. J. Int'l L. 495 (2008).

Lyons, Oren, Principles of Good Governance, Statement to the U.N. (May 15, 2014), https://www.onondaganation.org/news/2014/oren-lyons-presents-at-u-n-51514/.

Mačák, Kubo & Ewan Lawson, Avoiding Civilian Harm During Military Cyber Operations: Six Key Takeaways, ICRC Humanitarian L. & Pol'y Blog (June 15, 2021), https://blogs.icrc.org/law-and-policy/2021/06/15/avoiding-civilian-harm-military-cyber-operations/.

Mačák, Kubo, Tilman Rodenhäuser & Laurent Gisel, Cyber Attacks Against Hospitals and the COVID-19 Pandemic: How Strong Are International Law Protections?, ICRC Humanitarian L. & Pol'y Blog (Apr. 2, 2020), https://blogs.icrc.org/law-and-policy/2020/04/02/cyber-attacks-hospitals-covid-19/.

Maresca, L. & S. Maslen eds., The Banning of Anti-Personnel Landmines: The Legal Contribution of the ICRC (Cambridge Univ. Press 2000).

McDonald, A. & H. Brollowski, Requisitions, in R. Wolfrum ed., Max Planck Encyclopedia of Public International Law (Oxford Univ. Press, Apr. 2011).

McLachlan, Campbell, The Principle of Systemic Integration and Article 31(3)(c) of the Vienna Convention, 54 Int'l & Comp. L.Q. 279 (2005), https://www.wti.org/media/filer_public/88/b7/88b742e5-7030-4952-b38f-9dba323fde3b/iclq-mclachlan_principle_of_systemic_integration_and_article_313c_of_the_vienna_convention.pdf.

McNeill, J.H., Afterword: International Law and the Prohibition of Toxic Weapons, in A.H. Westing ed., Herbicides in War: The Long-Term Ecological and Human Consequences (Taylor & Francis 1984).

Meyrowitz, H., The Principle of Superfluous Injury or Unnecessary Suffering: From the Declaration of St. Petersburg of 1868 to Additional Protocol I of 1977, 34 Int'l Rev. Red Cross 98 (1994).

Microsoft Corporation, Microsoft Digital Defense Report 2024 (Oct. 2024), https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft%20Digital%20Defense%20Report%202024%20(1).pdf.

Military and Paramilitary Activities in and Against Nicaragua (Nicaragua v. United States of America), Merits, Judgment, 1986 I.C.J. Rep. 14 (June 27, 1986), https://www.icj-cij.org/case/70.

Moore, John Bassett, A Digest of International Law, vol. 2 (U.S. Government Printing Office 1906).

Murphy, V., Protection of the Natural Environment During Non-International Armed Conflict, in C. Stahn, J. Iverson & J.S. Easterday eds., Environmental Protection and Transitions from Conflict to Peace: Clarifying Norms, Principles, and Practices 220 (Oxford Univ. Press 2017).

Nahlik, S.E., Protection of Cultural Property, in UNESCO, International Dimensions of Humanitarian Law (Martinus Nijhoff 1988).

Nerlich, Volker. *Superior Responsibility Under Article 28 ICC Statute: For What Exactly Is the Superior Held Responsible?*, 5 J. Int'l Crim. Just. 665 (2007).

NSA, CISA, FBI et al., Joint Cybersecurity Advisory AA25-239A, Countering Chinese State-Sponsored Actors Compromise of Networks Worldwide to Feed Global Espionage Systems (Aug. 27, 2025, revised Sept. 3, 2025), https://www.cisa.gov/news-events/cybersecurity-advisories/aa25-239a.

Obregón Gieseken, H., The Protection of the Natural Environment Under International Humanitarian Law, in A.G. Khalifa ed., The Protection of the Environment During Armed Conflict: Proceedings of an Expert Meeting (UNEP 2009).

Omar, S.A.S., N.R. Bhat & A. Asem, Critical Assessment of the Environmental Consequences of the Invasion of Kuwait, the Gulf War, and the Aftermath, in T.A. Kassim & D. Barceló eds., Environmental Consequences of War and Aftermath 142 (Springer-Verlag 2009).

OPCW, Member States, https://www.opcw.org/about-us/member-states (last visited Apr. 16, 2026).

OPCW, Rules of Procedure of the Conference of the States Parties, Doc. C-I/3/Rev.2 (as amended), https://www.opcw.org/documents/2013/04/c-i3rev2/rules-procedure-conference-states-parties-organisation-prohibition.

OSCE, Vienna Document 2011 on Confidence- and Security-Building Measures, FSC.DOC/1/11 (Nov. 30, 2011), https://www.osce.org/files/f/documents/a/4/86597.pdf.

Parker, Arthur C., The Constitution of the Five Nations, New York State Museum Bulletin No. 184 (Apr. 1, 1916), https://archive.org/details/constitutionoffi00parkirich.

Pictet, J. ed., Commentary on the Fourth Geneva Convention (ICRC 1958).

Pictet, J., Development and Principles of International Humanitarian Law (Martinus Nijhoff/ICRC 1985).

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

Prosecutor v. Jean-Pierre Bemba Gombo, ICC-01/05-01/08, Judgment on the Appeal of Mr Jean-Pierre Bemba Gombo Against Trial Chamber III's "Judgment Pursuant to Article 74 of the Statute," ¶¶ 170–174 (Int'l Crim. Ct. June 8, 2018), https://www.icc-cpi.int/sites/default/files/CourtRecords/CR2018_02984.PDF.

Prosecutor v. Thomas Lubanga Dyilo, ICC-01/04-01/06, Trial Chamber I, Judgment Pursuant to Article 74 of the Statute (Int'l Crim. Ct. Mar. 14, 2012), https://www.icc-cpi.int/sites/default/files/CourtRecords/CR2012_03942.PDF.

Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), June 8, 1977, 1125 U.N.T.S. 3, https://ihl-databases.icrc.org/en/ihl-treaties/api-1977.

Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of Non-International Armed Conflicts (Protocol II), June 8, 1977, 1125 U.N.T.S. 609, https://ihl-databases.icrc.org/en/ihl-treaties/apii-1977.

Radics, G.B. & C. Bruch, The Law of Pillage, Conflict Resources, and Jus Post Bellum, in C. Stahn, J. Iverson & J.S. Easterday eds., Environmental Protection and Transitions from Conflict to Peace: Clarifying Norms, Principles, and Practices 143 (Oxford Univ. Press 2017).

Roberts, A. & R. Guelff eds., Documents on the Laws of War (3d ed., Oxford Univ. Press 2000).

Rogers, A.P.V., Law on the Battlefield (3d ed., Manchester Univ. Press 2012).

Rome Statute of the International Criminal Court, July 17, 1998, 2187 U.N.T.S. 90 (entered into force July 1, 2002), https://www.icc-cpi.int/sites/default/files/RS-Eng.pdf.

Sandoz, Yves, Christophe Swinarski & Bruno Zimmermann eds., Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949 (ICRC/Martinus Nijhoff 1987).

Sands, P. et al., Principles of International Environmental Law (4th ed., Cambridge Univ. Press 2018).

Sassòli, Marco, International Humanitarian Law: Rules, Controversies, and Solutions to Problems Arising in Warfare (2d ed., Edward Elgar 2024).

Schmitt, M.N. & J.J. Merriam, The Tyranny of Context: Isr. Targeting Practices in Legal Perspective, 37 U. Pa. J. Int'l L. 53 (2015).

Schmitt, M.N., C.H.B. Garraway & Y. Dinstein eds., The Manual on the Law of Non-International Armed Conflict, with Commentary (Institute of International Humanitarian Law, San Remo, 2006).

Schmitt, M.N., Green War: An Assessment of the Environmental Law of International Armed Conflict, 22 Yale J. Int'l L. 1 (1997).

Schmitt, Michael N. ed., Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations (2d ed., Cambridge Univ. Press 2017).

Schmitt, Michael N. & Jeffrey S. Thurnher, "Out of the Loop": Autonomous Weapon Systems and the Law of Armed Conflict, 4 Harv. Nat'l Sec. J. 231 (2013).

Silver, David et al., Mastering the Game of Go with Deep Neural Networks and Tree Search, 529 Nature 484 (2016).

Slack, Dylan, Sophie Hilgard, Emily Jia, Sameer Singh & Himabindu Lakkaraju, Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods, Proc. AAAI/ACM Conf. AI, Ethics & Soc'y 180 (2020), https://doi.org/10.1145/3375627.3375830.

Solis, G.D., The Law of Armed Conflict: International Humanitarian Law in War (2d ed., Cambridge Univ. Press 2016).

Spadaro, Alessandra, A Weapon Is No Subordinate: Autonomous Weapon Systems and the Scope of Superior Responsibility, 21 J. Int'l Crim. Just. 1119 (2023), https://doi.org/10.1093/jicj/mqad025.

Sparrow, Robert, Killer Robots, 24 J. Applied Phil. 62 (2007), https://robsparrow.com/wp-content/uploads/Killer-robots.pdf.

Special Representative of the Secretary-General, Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework, U.N. Doc. A/HRC/17/31 (Mar. 21, 2011), https://digitallibrary.un.org/record/706796.

Spoljaric, Mirjana, President, Int'l Comm. of the Red Cross, Risks from the Unconstrained Use of Autonomous Weapons in Armed Conflict Are Stark, Statement at the Luxembourg Autonomous Weapons Systems Conference (Apr. 26, 2023), https://www.icrc.org/en/document/risks-unconstrained-use-autonomous-weapons-armed-conflict-are-stark.

Stahn, C., J. Iverson & J.S. Easterday eds., Environmental Protection and Transitions from Conflict to Peace: Clarifying Norms, Principles, and Practices (Oxford Univ. Press 2017).

Statute of the International Atomic Energy Agency, Oct. 26, 1956, 276 U.N.T.S. 3 (entered into force July 29, 1957), https://www.iaea.org/about/statute.

Statute of the International Court of Justice, June 26, 1945, 59 Stat. 1031, T.S. No. 993, https://www.icj-cij.org/statute.

Sterio, Milena, Autonomous Weapons Systems and the Need to Update International Humanitarian Law?, 57 Case W. Res. J. Int'l L. 301 (2025), https://scholarlycommons.law.case.edu/cgi/viewcontent.cgi?article=2702&context=jil.

Stern, B. ed., Les aspects juridiques de la crise et de la guerre du Golfe (Montchrestien 1991).

Stroppa, Marta, The Use of Autonomous Cyber Capabilities in Armed Conflict and States' Due Diligence Obligations to Avoid Unintended Consequences, EJIL:Talk! (2023).

Szasz, Paul C., The Security Council Starts Legislating, 96 Am. J. Int'l L. 901 (2002).

Taghizadeh, Zakieh, Intergenerational and Intra-generational Equity Under the BBNJ Agreement, 75 Envtl. Mgmt. 3688 (2025), https://link.springer.com/article/10.1007/s00267-025-02256-5.

Testimony of Joseph Blount, President and Chief Executive Officer, Colonial Pipeline Company, Before the U.S. Senate Committee on Homeland Security and Governmental Affairs (June 8, 2021), https://www.hsgac.senate.gov/wp-content/uploads/imo/media/doc/Testimony-Blount-2021-06-08.pdf.

The CCP Cyber Threat to the American Homeland and National Security: Hearing Before the H. Select Comm. on Strategic Competition Between the U.S. & the Chinese Communist Party, 118th Cong. (Jan. 31, 2024), https://chinaselectcommittee.house.gov/sites/evo-subsites/selectcommitteeontheccp.house.gov/files/evo-media-document/1.31.24%20Hearing%20Transcript.pdf.

Trapp, Ralf & Cheng Tang, Enhancing the Management and Enforcement of Compliance in the Regime Prohibiting Chemical Weapons (UNIDIR 2021), https://doi.org/10.37559/WMD/21/CWC/01.

Treaty on the Non-Proliferation of Nuclear Weapons, July 1, 1968, 729 U.N.T.S. 161 (entered into force Mar. 5, 1970), https://www.un.org/disarmament/wmd/nuclear/npt/text.

Treaty on the Prohibition of Nuclear Weapons, July 7, 2017, U.N. Doc. A/CONF.229/2017/8 (entered into force Jan. 22, 2021), https://www.un.org/disarmament/wmd/nuclear/tpnw.

Tsagourias, Nicholas & Michael Farrell, Cyber Attribution: Technical and Legal Approaches and Challenges, 31 Eur. J. Int'l L. 941 (2020), https://academic.oup.com/ejil/article/31/3/941/5897247.

United Nations Security Council, Official Records, 72d Sess., 7922d mtg., *The Situation in the Middle East*, U.N. Doc. S/PV.7922 (Apr. 12, 2017), https://undocs.org/en/S/PV.7922.

United Nations Security Council, Official Records, 72d Sess., 8105th mtg., *The Situation in the Middle East*, U.N. Doc. S/PV.8105 (Nov. 16, 2017), https://undocs.org/en/S/PV.8105.

United Nations Security Council, France, United Kingdom of Great Britain and Northern Ireland and United States of America: Draft Resolution, U.N. Doc. S/2017/315 (Apr. 12, 2017), https://undocs.org/en/S/2017/315.

United Nations Security Council, France, Italy, Japan, Sweden, Ukraine, United Kingdom of Great Britain and Northern Ireland and United States of America: Draft Resolution, U.N. Doc. S/2017/962 (Nov. 16, 2017), https://undocs.org/en/S/2017/962.

U.K. Parliament, Public Accounts Committee, Cyber-Attack on the NHS: Oral Evidence, HC 787 (Feb. 5, 2018), https://committees.parliament.uk/work/6112/cyberattack-on-the-nhs/publications/.

U.N. Secretary-General, Lethal Autonomous Weapons Systems: Report of the Secretary-General, U.N. Doc. A/79/88 (May 2, 2024), https://documents.un.org/doc/un-doc/gen/n24/117/23/pdf/n2411723.pdf.

UN War Crimes Commission, History of the United Nations War Crimes Commission and the Development of the Laws of War (His Majesty's Stationery Office 1948).

UNEP, Environmental Rule of Law: First Global Report (UNEP 2019).

UNEP, From Conflict to Peacebuilding: The Role of Natural Resources and the Environment (UNEP 2009).

UNEP, Protecting the Environment During Armed Conflict: An Inventory and Analysis of International Law (UNEP 2009).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need." In *Advances in Neural Information Processing Systems*, vol. 31, 1–15. Curran Associates, 2017. https://arxiv.org/abs/1706.03762.

Vienna Convention on the Law of Treaties, May 23, 1969, 1155 U.N.T.S. 331 (entered into force Jan. 27, 1980), https://legal.un.org/ilc/texts/instruments/english/conventions/1_1_1969.pdf.

Von Glahn, G., The Occupation of Enemy Territory: A Commentary on the Law and Practice of Belligerent Occupation (Univ. of Minnesota Press 1967).

Watkin, K., Fighting at the Legal Boundaries: Controlling the Use of Force in Contemporary Conflict (Oxford Univ. Press 2016).

Westing, A.H. ed., Herbicides in War: The Long-Term Ecological and Human Consequences (Taylor & Francis 1984).

Westing, A.H., Ecological Consequences of the Second Indochina War (Stockholm International Peace Research Institute/Almqvist & Wiksell 1976).

White House, Fact Sheet: The Biden-Harris Administration Has Launched an All-of-Government Effort to Address Colonial Pipeline Incident (May 11, 2021), https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2021/05/11/fact-sheet-the-biden-harris-administration-has-launched-an-all-of-government-effort-to-address-colonial-pipeline-incident/.

White House, Statement from the Press Secretary Attributing NotPetya to Russian Military (Feb. 15, 2018), https://trumpwhitehouse.archives.gov/briefings-statements/statement-press-secretary-russia-destabilizing-cyber-activity/.

Wilmshurst, E. & S. Breau eds., Perspectives on the ICRC Study on Customary International Humanitarian Law (Cambridge Univ. Press 2007).

Zwanenburg, M., The Challenges of Applying the Principles of Distinction, Proportionality and Precautions in Contemporary Military Operations from a State Perspective, in ICRC/College of Europe, The Additional Protocols at 40: Achievements and Challenges, Proceedings of the 18th Bruges Colloquium, 19–20 October 2017, Collegium No. 48, at 151 (Autumn 2018).

Introduction

The Algorithmic Strike: Artificial Intelligence, Autonomous Cyber Weapons, and the Structural Incapacity of International Humanitarian Law

This dissertation advances one thesis: existing IHL is structurally incapable of governing autonomous cyber weapons — not because its rules are wrong, but because they presuppose a human being present at the moment of lethal targeting, and autonomous cyber weapons eliminate that human by design.

Three interlocking legal propositions together constitute this dissertation's central contribution to the international humanitarian law governing autonomous cyber weapons.

First, that the precautionary architecture of Additional Protocol I — specifically the verification obligation of Article 57(2)(a)(i), the cancellation obligation of Article 57(2)(b), and the weapons-review obligation of Article 36 — is not merely inadequately applied but structurally incapable of operational implementation when confronted with Tier 3 autonomous cyber weapons, whose four defining characteristics jointly generate a condition of "*verification impossibility*" that the drafters of the Protocol did not and could not have foreseen.

Second, that the attribution framework of the Articles on State Responsibility (ARSIWA) encounters a compliance deficit of categorical rather than remediable character, because the effective control standard presupposes a human decision-maker present at the moment of the operative act, a presupposition that Tier 3 architecture structurally defeats.

Third, that Article 28(1) of the Rome Statute, reoriented to the deployment authorization stage through the constructive knowledge

standard established in *Bemba Gombo*, supplies the legally coherent accountability locus that the first two deficiencies jointly require.

These three propositions are not parallel contributions that stand independently. They form a single argument: the first establishes the structural defect in IHL's compliance architecture; the second establishes that the same defect defeats accountability law; the third identifies the only legally coherent point at which accountability can be restored. Together they constitute the dissertation's thesis, the full implications of which are drawn in the Conclusion.

The urgency of these three propositions is demonstrated by a convergent operational record spanning three documented cyber operations from 2015 to 2024 and, most acutely, by the deployment of AI-assisted targeting systems in the United States-Iran conflict of February–April 2026. That record is analysed in Section I below. The normative framework whose structural deficiencies those operations expose is examined in Section II. The methodology and structure of the argument are set out in Sections III and IV respectively.

I. The Operational Record: Autonomous Cyber Execution and the Absence of Human Control

The legal crisis this dissertation examines did not arise in isolation. Operation Epic Fury — the deployment of AI-assisted targeting systems in the United States-Iran conflict of February–April 2026, analysed in Chapter 2, Section II below — is the most acute instantiation of a structural pattern observable across three prior documented operations spanning 2009 to 2016: autonomous cyber execution resulting in civilian harm without any human present, capable of intervening, or legally responsible at the operative moment of decision.

At approximately 3:35 p.m. local time on December 23, 2015, an operator at the Prykarpattyaoblenergo control center in western Ukraine

observed his cursor moving across the workstation screen without any movement of his hand, the visible signature of an adversary operating the system remotely. In the following thirty minutes, a coordinated cyber intrusion sequentially opened breakers at a minimum of twenty-seven distribution substations across three regional electricity companies, leaving approximately 225,000 civilians without electrical power.[1] The operator, present and fully aware that the attack was in progress, could not halt a single disconnection event; the malware had assumed control of the supervisory control and data acquisition interface and was executing a pre-programmed instruction sequence without further human input, rendering the human being with legal responsibility for the network structurally irrelevant by design.[2]

The Stuxnet weapon, the first autonomous cyber system documented to have achieved kinetic-equivalent physical destruction through software alone, demonstrated, between approximately June 2009 and mid-2010, that condition-based activation against pre-specified operational parameters and simultaneous false-status reporting to SCADA terminals could destroy approximately 1,000 uranium centrifuges at the Natanz Fuel Enrichment Plant. Iranian engineers observed falsified readings on their human-machine interface control screens throughout. No human

---

1 E-ISAC & SANS Institute, *Analysis of the Cyber Attack on the Ukrainian Power Grid* 3, 11 (Mar. 18, 2016) [hereinafter E-ISAC/SANS Ukraine Report] (confirming that adversaries used SCADA HMIs to open breakers at a minimum of twenty-seven substations across three oblenergos, impacting approximately 225,000 customers, and assessing with high confidence that the attack was highly synchronized and multisite); U.S. Dep't of Homeland Security, ICS-CERT, *Cyber-Attack Against Ukrainian Critical Infrastructure*, Alert IR-ALERT-H-16-056-01, at 1–2 (Feb. 25, 2016) [hereinafter ICS-CERT Ukraine Alert] (corroborating the 225,000-customer figure and confirming that attacks at each company occurred within thirty minutes of each other); *see also* Kim Zetter, *Inside the Cunning, Unprecedented Hack of Ukraine's Power Grid*, Wired (Mar. 3, 2016) (source for the 3:35 p.m. timestamp and operator-observed cursor-movement detail, drawn from contemporaneous company personnel accounts).

2 Robert M. Lee, Michael J. Assante & Tim Conway, *Analysis of the Cyber Attack on the Ukrainian Power Grid*, E-ISAC & SANS ICS Defense Use Case, at iv, 1, 6–8 (Mar. 18, 2016), https://media.kasperskycontenthub.com/wp-content/uploads/sites/43/2016/05/20081514/E-ISAC_SANS_Ukraine_DUC_5.pdf (reporting "seven 110 kV and 23 35 kV substations disconnected" at Prykarpattyaoblenergo "at approximately 3:35 p.m. local time," with adversary access established "more than six months prior to December 23, 2015"); NCCIC/ICS-CERT, Alert IR-ALERT-H-16-043-01AP, *Cyber-Attack Against Ukrainian Critical Infrastructure (Update A)*, at 1, 3–5 (U.S. Dep't of Homeland Security Mar. 7, 2016), https://info.publicintelligence.net/NCCIC-UkrainianPowerAttack.pdf (confirming "local operators were locked out of their own workstations, disabling keyboard and mouse control," and that spear-phishing commenced "as early as March 2015").

being was present, capable of intervening, or legally responsible at the operative moment of activation.[3]

The Industroyer/CRASHOVERRIDE malware, deployed against Ukraine's transmission-level electricity grid on December 17, 2016, incorporated autonomous device enumeration and, following a pre-programmed activation timestamp, executed all payloads without any real-time human command, a design independently verified by ESET and Dragos in simultaneous forensic analyses of June 12, 2017, confirming that the IEC 104 module entered an infinite loop of switch-State commands overriding all operator input, that real-time human intervention was operationally impossible during execution.[4]

Over a span of seven years, three operations — attributed with varying degrees of confidence to Western State actors (Stuxnet),[5] with high

3 Ralph Langner, *Stuxnet: Dissecting a Cyberwarfare Weapon*, IEEE Security & Privacy, vol. 9, no. 3, May–June 2011, at 49, 50–51 ("the 417 attack code . . . provided the legitimate program code with a fake process image," feeding "prerecorded input to the legitimate code during the attack" so that operators observed falsified readings on their human-machine interface control screens throughout; "the attack wasn't remotely controlled—it was completely stand-alone and didn't require Internet access"; attacks "were triggered by complex timer and process conditions" without any real-time human command), https://doi.org/10.1109/MSP.2011.67 (last visited May 21, 2026); Int'l Atomic Energy Agency, *Implementation of the NPT Safeguards Agreement and Relevant Provisions of Security Council Resolutions in the Islamic Republic of Iran*, IAEA Doc. GOV/2010/62, ¶ 4 & n.3 (Nov. 23, 2010) (reporting detection of a "new generation" of centrifuges at Natanz and confirming that 328 of 1,064 IR-1 centrifuges installed in Unit A of Fuel Enrichment Plant cascades had been replaced, consistent with physical damage during operation) [hereinafter IAEA Doc. GOV/2010/62]; David Albright, Paul Brannan & Christina Walrond, *Did Stuxnet Take Out 1,000 Centrifuges at the Natanz Enrichment Plant?*, ISIS Report, at 1–3, 8–9 (Inst. for Sci. & Int'l Sec., Dec. 22, 2010) (documenting the reduction from approximately 4,920 to 3,936 IR-1 centrifuges operating at Natanz between May and August 2009, and inferring, through cross-referencing IAEA safeguards data, physical destruction of approximately one thousand centrifuges attributable to the Stuxnet payload), https://isis-online.org/uploads/isis-reports/documents/stuxnet_FEP_22Dec2010.pdf (last visited May 21, 2026) [hereinafter Albright, Brannan & Walrond, ISIS Report].

4 Anton Cherepanov, *WIN32/INDUSTROYER: A New Threat for Industrial Control Systems*, ESET White Paper, at 5, 8–9, 15 (June 12, 2017), https://www.welivesecurity.com/wp-content/uploads/2017/06/Win32_Industroyer.pdf (last visited May 21, 2026) (documenting that the Launcher component contained a hardcoded activation timestamp of "17th December 2016, the day of the power outage," upon which it "creates two threads" executing all payloads autonomously, with the 104 payload thereafter executing "an infinite loop" of switch-state commands without further human input); Dragos, Inc., *CRASHOVERRIDE: Analyzing the Threat to Electric Grid Operations*, at 16, 23 (June 12, 2017), https://www.dragos.com/wp-content/uploads/CrashOverride-01.pdf (last visited May 21, 2026) ("control then passes from the launcher to the loaded module" without further operator command; "the command then begins an infinite loop and continues to set addresses to this value effectively opening closed breakers"; "if a system operator tries to issue a close command on their HMI the sequence loop will continue to re-open the breaker," rendering real-time human intervention operationally impossible during execution).

5 Attribution of Stuxnet to specific States remains formally unacknowledged by those States; the United States government has neither confirmed nor denied responsibility. The attribution to Western State actors, widely accepted in the technical and intelligence literature, is relied upon here as the prevailing analytical consensus, not as an authoritative legal finding.

confidence to Russian State-sponsored actors (BlackEnergy/Sandworm and Industroyer/CRASHOVERRIDE) — demonstrated with operational precision one consistent design logic: autonomous cyber weapons producing civilian-affecting physical effects without any human presence, capacity for intervention, or legal responsibility at the operative moment of activation.[6] That is not a gap in compliance. It is the elimination of the factual precondition upon which IHL's compliance architecture depends.

II. Three Structural Deficiencies of the Existing Normative Framework

The operational record in Section I demonstrates, in concrete terms, the three structural deficiencies that the three propositions above identify in doctrinal terms.

First, a definitional deficiency: no binding international legal instrument defines an autonomous cyber weapon with sufficient precision to ground a legal compliance assessment.

Second, a precautionary deficiency: IHL's precautionary framework under AP I cannot operate for Tier 3 systems because its obligations presuppose a human decision-maker at the moment of targeting — a human these weapons eliminate by design.

Third, an accountability deficiency: ARSIWA's attribution framework and the Rome Statute's criminal responsibility provisions both presuppose a human being in the causal chain, a presupposition Tier 3 architecture structurally defeats. Each deficiency is examined and proven in the chapters that follow.

The three structural deficiencies introduced in Section II above generate three research questions that structure the dissertation's analytical sequence.

6 Cherepanov, *supra* note 4, at 2, 5, 11–12, 15; Dragos, Inc., *CRASHOVERRIDE: supra* note 4, at 3–4, 12.

Research Question 1 (Definitional): How can autonomous cyber weapons be defined and classified under international law, what degrees of autonomy exist across the Tier 1–3 spectrum, and what distinct legal implications flow from each tier of autonomous operation?

Research Question 2 (IHL Application): Are the existing IHL principles of distinction, proportionality, and precaution under AP I Articles 48, 51(5)(b), 52(3), and 57 capable of operational implementation when confronted with the architecture of Tier 3 autonomous cyber weapons, and if not, does that incapacity derive from a remediable gap in State practice or from a structural feature of Tier 3 systems that no amendment to practice can cure?

Research Question 3 (Governance): What preventive governance architecture is required to address ACW's specific challenges for human security, and through what institutional mechanisms, including a proposed International Autonomous Weapons Control Organization, can that architecture be anchored, verified, and enforced?

III. Methodology, Sources, and Epistemic Framework

The dissertation operates within the positivist tradition, identifying binding obligations exclusively through Article 38(1) ICJ Statute sources: treaty, custom, general principles, and the judicial decisions and scholarly writings that evidence them. A strict distinction between *lex lata* findings (Chapters 1–4) and *lex ferenda* proposals (Chapters 5–6) is maintained throughout; where the argument crosses from existing law to proposed law, this is explicitly marked.

The three documented operational cases — Stuxnet, the 2015 BlackEnergy/KillDisk operation against Ukraine's electricity distribution network, and Industroyer/CRASHOVERRIDE — are analysed as publicly documented technical and operational records, not as legal findings of State responsibility; the factual premises drawn from those records are stated as the evidentiary basis for the legal analysis,

not as adjudicated facts. Operation Epic Fury is analysed on the assumption that the publicly reported facts are accurate and that the deployed systems satisfy the Tier 3 criteria established in Chapter 1 — an epistemic assumption stated explicitly wherever that analysis is engaged.

IV. Structure of the Argument

Chapter 1 establishes the technological and definitional foundations of the inquiry: the three-tier classification system, the four-characteristic analytical framework and the attribution-resistance aggravating factor, the dual-use supply chain problem, and the systemic integration methodology grounded in Vienna Convention on the Law of Treaties Article 31(3)(c).

Chapter 2 demonstrates that the *jus contra bellum* — the Article 2(4) prohibition, the Article 51 self-defense right, the Caroline conditions of necessity and imminence, and the ARSIWA countermeasures regime — cannot accommodate the temporal disaggregation between deployment authorization and autonomous activation that Tier 3 architecture generates by design. The pre-authorization deficit established in Chapter 2 is not a gap in State practice. It is a structural feature of the weapon.

Chapter 3 demonstrates that the same structural defect defeats the precautionary architecture of Additional Protocol I across distinction, proportionality, precaution, and the special protections for medical units, objects indispensable to civilian survival, and works containing dangerous forces. The Algorithmic Transparency Obligation — the dissertation's principal original doctrinal contribution — is established in Chapter 3 as what existing law already requires.

Chapter 4 traces the accountability consequences of the structural incompatibility established in Chapters 2 and 3: command responsibility reoriented to the deployment authorization stage through the *Bemba Gombo* constructive knowledge standard, the ARSIWA attribution gap,

and the Rome Statute's Article 28 as the legally coherent locus of accountability.

Chapter 5 presents the Human Security by Design framework: the four technical design criteria — kill-switch override, geofencing, temporal deactivation logic, and tamper-proof audit logging — that translate IHL's existing obligations into requirements an engineer can build and a lawyer can verify.

Chapter 6 develops the IAWCO governance architecture, grounding its political feasibility in the majority expressed in UNGA Resolution 78/241, the Secretary-General's 2024 report, and the accumulated record of the CCW GGE process.

Annex I presents a model IAWCO Convention with Article-by-Article negotiating commentary, operationalizing the normative and institutional proposals advanced in Chapters 5 and 6.

Two boundaries define the scope of the dissertation's central claim — that existing IHL is structurally incapable of governing Tier 3 autonomous cyber weapons.

First, it applies only to Tier 3 systems — systems whose targeting decisions are generated by machine learning without real-time human verification. Tier 1 and Tier 2 systems, which retain meaningful human control at the target-selection node, fall outside it.

Second, the claim is *lex lata*: it identifies what existing law already requires and already prohibits, not what new law should say. The governance architecture of Chapters 5 and 6 follows from the *lex lata* diagnosis; it is the institutional response the existing law demands, not a replacement for it.

These boundaries together constitute the argument of the six chapters that follow.

The red line running through all six chapters is this: IHL presupposes that a human being is present at the moment of lethal targeting. Autonomous cyber weapons eliminate that human by design. Every chapter is one step in demonstrating what that elimination means — legally, structurally, and institutionally.

Chapter 1 defines the weapon that removes the human.

Chapter 2 demonstrates what the *jus contra bellum* cannot accommodate when that human is absent at the moment of force.

Chapter 3 demonstrates what the *jus in bello* under AP I cannot accommodate when that same human is absent at the moment of attack.

Chapter 4 demonstrates the accountability consequences of that dual structural failure.

Chapter 5 provides the normative answer: Human Security by Design.

Chapter 6 builds the institution that makes that answer enforceable in practice: the IAWCO.

Chapter 1
Autonomous Cyber Weapons: Technological Architecture and Legal Classification

International humanitarian law governs weapons. It does not govern abstractions. Before any legal claim about autonomous cyber weapons can be assessed — before a precautionary obligation can be applied, a prohibition established, or a command responsibility determined — the technology itself must be defined with sufficient precision to serve as the subject of legal analysis. That precision is what this Chapter supplies.

Chapter 1 establishes one thing: what a Tier 3 autonomous cyber weapon is. It does so through two analytical moves. First, it identifies the four technological characteristics that distinguish Tier 3 systems from all prior means of warfare in legally determinative ways — propagation dynamics, temporal persistence, technical irreversibility, and attribution opacity.

Second, it establishes the three-tier functional taxonomy — human-in, human-on, and human-out-of-the-loop — and demonstrates why Tier 3 systems, defined by the structural elimination of human judgment at the moment of target selection, occupy a categorically different legal position from the tiers below them.

Every legal claim in the chapters that follow rests on the definitional conclusions this Chapter establishes. A reader who accepts those conclusions will find that the structural incompatibility thesis of Chapters 2 and 3, the accountability analysis of Chapter 4, and the governance architecture of Chapters 5 and 6 each follow necessarily from applying existing law to them. A reader who contests those conclusions should do so here — they are the empirical predicate upon which the entire normative argument rests. The sustained legal argument

begins in Chapter 2, which asks whether the law governing resort to force — the *jus contra bellum* — can accommodate a weapon whose human has been eliminated before force arrives.

I. Definitional Framework and Technological Parameters

The distinction between automated and autonomous weapons systems is legally determinative. An automated system executes pre-programmed instructions in response to environmental inputs without altering its decision parameters; an autonomous system selects targets and applies force through machine learning algorithms[7] that adapt behavior based on operational feedback. International humanitarian law's precautionary obligations presuppose that a human decision-maker evaluates each specific attack against the requirements of distinction, proportionality, and precaution at the moment force is applied. Automated systems preserve that human evaluation; autonomous systems eliminate it.

This distinction generates the structural incompatibility thesis that this dissertation advances. Where an autonomous weapon system's targeting parameters evolve through machine learning after deployment, no human operator can verify compliance with Article 57(2)(a)(i)–(iii) of Additional Protocol I[8] for the specific engagement the system executes.

1. The Automation–Autonomy Continuum and IHL

International governance frameworks recognize the automation-autonomy distinction but have not translated it into binding legal obligations. The CCW Group of Governmental Experts' Rolling Text, in its November 2024 version, provisionally characterizes a lethal

7 U.S. Dep't of Defense, DoD Directive 3000.09, *Autonomy in Weapon Systems* §§ 1.2(a)–(c), G.2 (Jan. 25, 2023); Vincent Boulanin et al., Artificial Intelligence, Strategic Stability and Nuclear Risk: Euro-Atlantic Perspectives 15–18 (Stockholm Int'l Peace Research Inst. 2020) (distinguishing automated systems that "logically follow a pre-defined set of rules" from autonomous systems that can "independently compose and select among different courses of action" and "execute a task or tasks using explicit or implicit programming rules and without human intervention").

8 AP I, art. 57(2)(a)(i)–(iii).

autonomous weapon system as one whose technological components "enable the system to identify and/or select, and engage a target, without intervention by a human user in the execution of these tasks," while the ICRC's working definition, as reproduced in the UN Secretary-General's July 2024 report, describes autonomous weapons systems as those "designed to select and engage one or more targets without the need for human intervention after activation."[9] Neither instrument establishes the legal threshold at which autonomy becomes incompatible with IHL's precautionary obligations, the threshold this subsection identifies.

The automation-autonomy continuum ranges from purely mechanical systems executing fixed instruction sets to learning-based algorithms that adapt targeting parameters in response to operational feedback. At the lower end, signature-based intrusion detection systems apply hard-coded matching logic to known threat patterns.[10] At the upper end, reinforcement learning agents optimize behavior through trial-and-error interaction with operational environments, adapting strategies based on reward signals without human intervention.[11] The legally determinative threshold lies where machine learning eliminates ex ante verification of targeting decisions, the point at which commanders cannot meaningfully predict which specific targets will be struck, when

9 CCW Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Rolling Text (version circulated Nov. 8, 2024), Box I, ¶ 1 [hereinafter CCW GGE Rolling Text (Nov. 2024)] (informal document without formal U.N. symbol, circulated intersessionally; confirmed in Chair's Summary - First 2025 Session of the GGE on LAWS, U.N. Doc. CCW/GGE.1/2025/WP.1, ¶ 9 (Apr. 7, 2025)); U.N. Secretary-General, *Lethal Autonomous Weapons Systems: Report of the Secretary-General*, ¶ 8, U.N. Doc. A/79/88 (July 1, 2024) [hereinafter SG LAWS Report 2024] (reproducing ICRC working definition). Note that the May 2025 revision of the Rolling Text replaced "identify and/or select, and engage" with "select and engage," on the Chairperson's understanding that identification is a precondition of selection; *see* Chair's Summary - Second 2025 Session of the GGE on LAWS, U.N. Doc. CCW/GGE.1/2025/WP.9, ¶ 10 (Oct. 20, 2025). The present analysis is grounded in the November 2024 text as the last version preceding the dissertation's primary research cut-off, with subsequent doctrinal developments noted where material.

10 Jesús Díaz-Verdejo et al., *On the Detection Capabilities of Signature-Based Intrusion Detection Systems in the Context of Web Attacks*, 12 Appl. Sci. 852, 852–53 (2022).

11 A. Manerba, L. Pajola, F. Donadel & M. Conti, *Demystifying the Role of Rule-Based Detection in AI Systems for Windows Malware Detection*, in Proceedings of the IEEE Symposium on Security & Privacy 1, 1–2 (2025); Sanyam Vyas et al., *Automated Cyber Defense: A Review*, arXiv:2303.04926, at 1–2, 6 (working paper 2023).

engagements will occur, or whether individual attacks will comply with Articles 48, 51(5)(b), and 57 of AP I.[12]

2. Four Technological Characteristics Distinguishing Autonomous Cyber Weapons

Four characteristics distinguish Tier 3 autonomous cyber weapons from all prior means of warfare in legally determinative ways: (a) propagation dynamics that defeat spatial boundedness; (b) temporal persistence and dormancy that defeat the contemporaneity of precautionary assessment; (c) technical irreversibility that defeats the Article 57(2)(b) AP I cancellation obligation; and (d) attribution opacity that compounds the structural accountability deficit. Each is analyzed below exclusively within the lex lata framework of AP I, ARSIWA, and the UN Charter.

a) Propagation Dynamics

Kinetic autonomous weapon systems operate within bounded spatial envelopes determined by physical constraints - fuel capacity, propulsion range, munitions payload - that confine their destructive effects to the geographic area within which the deploying commander authorized engagement. ACW, by contrast, violate these bounded operational assumptions through self–propagation capabilities that enable systems to traverse network interconnections, discover vulnerabilities in adjacent infrastructure, and replicate across digital systems far beyond initial deployment parameters. The 2017 NotPetya operation illustrates this propagation dynamic with operational precision: initially deployed against Ukrainian financial institutions, it propagated globally within hours, causing losses estimated at ten billion U.S. dollars to entities remote from any military objective, a pattern analyzed in detail at

12 Int'l Comm. of the Red Cross, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts* 2, 6 (2019) [hereinafter ICRC IHL Challenges Report 2019]; Neil Davison, *A Legal Perspective: Autonomous Weapon Systems Under International Humanitarian Law*, *in* UNODA Occasional Papers No. 30, at 9, 9–11, 15 (2017). A further research avenue concerns the intersectional dimensions of algorithmic warfare, specifically, the documented evidence that automated targeting systems trained on historical conflict data replicate structural patterns of discrimination against civilian populations, requiring systematic engagement with algorithmic bias scholarship as a distinct axis of IHL compliance analysis; this question is reserved for separate treatment.

Chapter 2, Section IV.[13] The crucial point for IHL is not the precise quantum of financial loss, but the demonstration that effects can manifest in States not party to the conflict and in sectors with no discernible connection to the original military objective. From the perspective of an attacker seeking to comply with the requirement to limit attacks to military objectives, the capacity of code to move through complex, globally interconnected networks introduces a degree of spatial indeterminacy unknown in the kinetic paradigm. Even if initial target selection meets all applicable criteria, the technical behavior of self–propagating malware may carry the operation far beyond the area that could reasonably have been contemplated at the point of authorization.[14]

b) Temporal Persistence and Dormancy

Conventional kinetic AWS exhibit temporal immediacy; once deployed and authorized, they engage targets within operationally relevant timeframes, typically minutes to hours, enabling commanders to assess battlefield conditions contemporaneously with engagement. This temporal proximity between authorization and effect permits meaningful human judgment over the use of force and facilitates compliance with obligations to verify, at the time of attack, that objectives are military objectives, and that expected incidental harm is

13 Press Secretary, *Statement from the Press Secretary*, The White House (Feb. 15, 2018), https://trumpwhitehouse.archives.gov/briefings-statements/statement-press-secretary-25/ (last visited Apr. 16, 2026); *Foreign Office Minister Condemns Russia for NotPetya Attacks*, U.K. Foreign & Commonwealth Off. (Feb. 14, 2018), https://www.gov.uk/government/news/foreign-office-minister-condemns-russia-for-notpetya-attacks (last visited Apr. 16, 2026); *This Is the Costliest Cyber Attack in History*, Geneva Ctr. for Sec. Policy (Aug. 21, 2019), https://www.gcsp.ch/global-insights/costliest-cyber-attack-history (last visited Apr. 16, 2026); Warwick Ashford, *NotPetya Attack Cost up to $300m, Says Maersk*, Computer Wkly. (Aug. 15, 2017); *White House Blasts Russia for NotPetya Cyberattack*, CNN Politics (Feb. 15, 2018), https://www.cnn.com/2018/02/15/politics/white-house-russia-notpetya (last visited Apr. 16, 2026).

14 Int'l Comm. of the Red Cross, *The Potential Human Cost of Cyber Operations* 38–39 (Nov. 2018), https://www.icrc.org/en/download/file/97346/the-potential-human-cost-of-cyber-operations.pdf (last visited Apr. 16, 2026) [hereinafter ICRC, *Potential Human Cost*]; Int'l Comm. of the Red Cross, *International Humanitarian Law and Cyber Operations During Armed Conflicts*, 102 Int'l Rev. Red Cross 481, at 5 (2020) [hereinafter ICRC Cyber Operations Report].; Lockheed Martin, *Gaining the Advantage: Applying Cyber Kill Chain® Methodology to Network Defense*, at phases 3–5 (Delivery, Exploitation & Installation) (2015), https://www.lockheedmartin.com/content/dam/lockheed-martin/rms/documents/cyber/Gaining_the_Advantage_Cyber_Kill_Chain.pdf (last visited Apr. 16, 2026) [hereinafter Lockheed Martin, *Cyber Kill Chain*].

not excessive in relation to the anticipated military advantage.[15] Commanders and legal advisers can, at least in principle, align their assessment with current intelligence and adapt plans if the situation on the ground changes in ways that materially affect the proportionality calculus.

Temporal immediacy thereby maintains the causal proximity between situational assessment and operational execution upon which the operationalization of both the Article 57(2)(a)(i) AP I verification obligation and the Article 57(2)(b) AP I cancellation obligation jointly depend. It also underpins the expectation, reflected in doctrine and training, that attacks can be cancelled or postponed up to the last moment if new information emerges.[16]

ACW can remain dormant within compromised systems for extended periods, months, or even years, before activating, thereby creating temporal displacement between authorization and effects that frustrates commanders' ability to fulfill ongoing precautionary obligations. Advanced persistent threat architectures establish covert access, monitor for predefined trigger conditions, and activate offensive capabilities only when those conditions are met, potentially long after the initial deployment decision. During such dormancy periods, the technical environment into which the code was introduced may change in ways that were neither foreseen nor foreseeable at the time of authorization. Civilian infrastructure dependencies may evolve in ways that render an operation which was initially assessed as proportionate manifestly disproportionate under changed circumstances; political contexts may transform, rendering initially lawful military objectives protected under subsequent agreements or altering the legal characterization of the situation; yet commanders may lack any practical

15 AP I, art. 57(2)(a); ICRC, Int'l Comm. of the Red Cross, *Autonomous Weapon Systems and International Humanitarian Law: Selected Issues* at 5, 8 (Mar. 2026), https://www.icrc.org/en/Article/autonomous-weapon-systems-and-international-humanitarian-law-selected-issues (last visited Apr. 16, 2026) [hereinafter ICRC, *AWS Selected Issues*].

16 AP I, art. 57(2)(b); ICRC, *AWS Selected Issues*, *supra* note 15, at 16.

technical capability to reassess proportionality or cancel the operation before dormant systems activate autonomously.[17]

c) Technical Irreversibility

Conventional kinetic AWS generally permit a meaningful degree of reversibility. Communications links between autonomous platforms and human operators, as well as built-in abort functions or self-destruct mechanisms, typically allow missions to be terminated when battlefield conditions change or when new information indicates that the proportionality or distinction assessment has shifted. Kinetic autonomous platforms thus preserve, at least architecturally, the human intervention capacity that Article 57(2)(b) AP I requires, the capacity to cancel or suspend an attack when its unlawfulness becomes apparent.[18] Tier 3 autonomous cyber weapons, by contrast, structurally foreclose this capacity: once self-propagating code is introduced into a network environment, the technical conditions for cancellation or suspension, continued operator access, identifiable system State, and controllable execution cannot be assumed to persist beyond the moment of deployment.[19]

ACW deployed across distributed network infrastructures cannot be reliably recalled or deactivated once propagation has occurred beyond the deploying State's technical control. Malware replicating across multiple systems-potentially traversing international cloud service providers operating under diverse jurisdictions or commercial telecommunications networks serving civilian populations-creates

17 Int'l Comm. of the Red Cross, *The Potential Human Cost of Cyber Operations* at 39 (Nov. 2018), https://www.icrc.org/en/download/file/97346/the-potential-human-cost-of-cyber-operations.pdf (last visited Apr. 16, 2026) [hereinafter ICRC, *Potential Human Cost*]; ICRC, *AWS Selected Issues*, *supra* note 15, at 16.

18 AP I, art. 57(2)(b); Yves Sandoz, Christophe Swinarski & Bruno Zimmermann eds., COMMENTARY ON THE ADDITIONAL PROTOCOLS OF 8 JUNE 1977 TO THE GENEVA CONVENTIONS OF 12 AUGUST 1949 ¶ 2220 (ICRC & Martinus Nijhoff 1987) [hereinafter 1987 ICRC AP Commentary]; Michael N. Schmitt ed., Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations r. 119 cmt. ¶¶ 5, 7 (Cambridge Univ. Press 2d ed. 2017) [hereinafter Tallinn Manual 2.0].

19 Int'l Comm. of the Red Cross, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts* 1425–26 (2024) [hereinafter ICRC IHL Challenges Report 2024].

situations in which the deploying State no longer possesses either practical access to or exclusive technical control over the code it has released. Even where technical "kill switches" or update mechanisms are embedded in the initial design, their effectiveness depends on continued connectivity, cooperation by intermediaries, and the absence of countermeasures by adversaries or third parties. In some cases, attempts to halt or delete malicious code may themselves cause further disruption, for example, by triggering defensive responses in compromised systems or by destabilizing fragile networks. Even if political authorities subsequently determine that continuation of the operation would be incompatible with proportionality or distinction obligations, they may therefore lack any realistic means to suspend the effects already unfolding in third–party systems.[20]

d) Attribution Opacity and the Structural Accountability Deficit

Conventional kinetic AWS maintain relatively straightforward attribution chains. States procure weapons through identifiable defense contractors, exercise operational command through recognized military hierarchies, and deploy platforms bearing national markings or operating from territory or bases clearly under their control. These features facilitate the application of the general law of State responsibility, which requires that conduct be attributable to a State in order to constitute an internationally wrongful act.[21] Factfinders can often rely on physical remnants, radar tracks, eyewitness testimony, and

---

20 Int'l Comm. of the Red Cross, *The Potential Human Cost of Cyber Operations* 38–39 (Nov. 2018), https://www.icrc.org/en/download/file/97346/the-potential-human-cost-of-cyber-operations.pdf (last visited Apr. 16, 2026) [hereinafter ICRC, *Potential Human Cost*]; Int'l Comm. of the Red Cross, *International Humanitarian Law and Cyber Operations During Armed Conflicts*, 102 Int'l Rev. Red Cross 481, at 4–6 (2020) [hereinafter ICRC Cyber Operations Report].; Monica Kaminska, Fabio Cristiano & Dennis Broeders, *Limiting Viral Spread: Automated Cyber Operations and the Principles of Distinction and Discrimination in the Grey Zone*, *in* 13th International Conference on Cyber Conflict: Going Viral 60, 61–62 (NATO CCDCOE 2021), https://ssrn.com/abstract=3856623 (last visited Apr. 16, 2026).

21 Int'l Law Comm'n, Responsibility of States for Internationally Wrongful Acts, U.N. Doc. A/RES/56/83, arts. 2, 4, 8 annex (Dec. 12, 2001) [hereinafter ARSIWA] https://legal.un.org/ilc/texts/instruments/english/draft_articles/9_6_2001.pdf (last visited Apr. 16, 2026); Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosn. & Herz. v. Serb. & Montenegro), Merits, Judgment, 2007 I.C.J. Rep. 43, ¶¶ 385, 388, 396–97 (Feb. 26), https://www.icj-cij.org/case/91 (last visited Apr. 16, 2026).

other traditional forms of evidence to link specific attacks to specific armed forces. The institutional embeddedness of kinetic operations within formal military structures thus supports the functioning of accountability mechanisms at both the international and domestic levels.[22]

ACW exhibit attribution complexity arising from three structural characteristics that complicate the application of existing State responsibility frameworks. First, dual–use components, whereby weapons leverage commercial artificial intelligence platforms-cloud–based machine learning services, open–source neural network libraries, commercial threat intelligence feeds-are functionally and technically indistinguishable from legitimate cybersecurity tools, thereby impeding attribution through technical component analysis or export–control–based monitoring. Second, distributed development chains, in which machine learning models may be trained by commercial technology firms, deployed via cloud infrastructure operated by multinational corporations incorporated in neutral States, and controlled through command servers hosted by third–party providers, fragment causal responsibility across multiple non–State actors whose conduct is not readily attributable to States under the restrictive standards governing control over private actors.[23] Third, algorithmic opacity - where complex machine learning systems' decision pathways are not intelligible even to their designers - limits the usefulness of behavioral analysis for evidentiary purposes, thereby increasing the difficulty of establishing, to the requisite standard of

22 ARSIWA, *supra* note 21, art. 2; Eurojust, *Eurojust Memorandum on Battlefield Evidence* 5, 8–9, 11–13 (Sept. 2020), https://www.eurojust.europa.eu/sites/default/files/assets/2020_09_14_eurojust_memorandum_on_battlefield_evidence.pdf (last visited Apr. 16, 2026).

23 Katharina Ziolkowski, *The Attribution of Cyber Operations to States in International Law* 12–13, 15 (Geneva Ctr. for Sec. Policy 2025), https://www.gcsp.ch/publications/attribution-cyber-operations-states-international-law (last visited Apr. 16, 2026) [hereinafter Ziolkowski, *Attribution*].

proof, that particular cyber effects are the result of a given State's autonomous cyber operation.[24]

3. The Verification Incapacity of Machine–Learning Architectures

Machine-learning architectures generate a fourth structural obstacle: verification incapacity. Deep neural networks produce classification outputs through mathematical transformations that their own designers cannot fully explain. A commander presented with an autonomous weapon's targeting output — "military objective: 87% confidence" — cannot verify whether that output reflects a legally sound application of Article 52(2) AP I or an artifact of training-data correlation. The opacity is not a temporary engineering limitation. It is a constitutive feature of the deep-learning paradigm. This dissertation's analysis of its legal consequences is developed in Chapter 3, Section II.

II. Functional Classification of Autonomous Weapon Systems: Doctrinal Architecture and Its Limits

Western military doctrine defines autonomous weapon systems functionally, through criteria emphasizing operational capabilities rather than technological implementation. The three–tier classification that follows draws on that functional tradition, adapting it to the specific compliance demands that Article 48, 51(5)(b), and 57 AP I impose on targeting systems in the cyber operational context.

Tier 3 autonomous weapon systems, once activated, can select and engage targets without further intervention by a human operator, though current U.S. policy explicitly excludes autonomous cyberspace capabilities from this doctrinal framework.[25] Target identification in autonomous systems employs machine learning algorithms to analyze

---

24 Nicholas Tsagourias & Michael Farrell, *Cyber Attribution: Technical and Legal Approaches and Challenges*, 31 Eur. J. Int'l L. 941, 956–58 (2020), https://academic.oup.com/ejil/article/31/3/941/5897247; ICRC IHL Challenges Report 2024, *supra* note 19, at 65.

25 DoD Directive 3000.09, *supra* note 7 §§ 1.1(b)(1).

operational data and distinguish military objectives from civilian objects. An expert workshop convened by Kwik, Zwanenburg, Took, and Aarts found that IHL specialists were more willing to authorize AI-enabled targeting systems operating at a false positive rate of 2.5% or lower, while rates of 7.5% and above generated significant reservations; the authors caution against treating this finding as a definitive legal threshold, given the workshop's limited sample size and the inseparability of legal and policy considerations in expert deliberations.[26] These findings are cited here as empirical evidence of the domain in which Article 48 AP I compliance concerns arise, not as establishing a quantitative legal threshold, which the existing law does not supply. Target selection involves algorithmic processes weighing military utility, operational vulnerability, and the anticipated effects' scope to prioritize engagement sequences when multiple candidates present tactical opportunities. Target engagement executes through automated deployment of cyber effects based on algorithmic decisions rather than human authorization for each individual action.[27]

The critical distinction between autonomous and semi–autonomous weapon systems lies in the temporal dimension of human intervention: semi–autonomous systems require human operators to select specific targets before engagement, while autonomous systems select and engage targets without further human intervention after initial activation, creating fundamental uncertainty regarding which specific targets will be attacked and when.[28] This temporal dimension proves

26 Jonathan Kwik, Marten Zwanenburg, Cheya Took & Joep Aarts, *Controlling Military Artificial Intelligence: Harnessing Rules of Engagement and Military Directives*, Asser Policy Brief 2025-01, at 19–20 (Asser Inst. Feb. 2025) [hereinafter Kwik et al., *Controlling Military AI*] (reporting that a workshop of IHL experts "appeared more willing to authorise the system under conditions corresponding to a 2.5% false positive rate," while cautioning that "caution is advised against drawing definitive conclusions about accuracy 'cut-off points' based solely on these results" given the n = 2 sample size and the difficulty of disaggregating legal from policy considerations in operational expert deliberations; a 20% false positive rate was unanimously rejected as incompatible with AP I art. 51(4); no authoritative IHL instrument has established any numerical threshold as a binding legal standard).

27 AP I, art. 57(2)(a)(ii); Int'l Comm. of the Red Cross, *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons* 8–9, 12–13 (2016) [hereinafter ICRC AWS 2016 Report]; Kwik et al., *Controlling Military AI*, *supra* note 26, at 19–20.

28 ICRC AWS 2016 Report, *supra* note 27, at 8; Kwik et al., *Controlling Military AI*, *supra* note 26, at 19–20.

legally determinative because Article 57 AP I precautionary obligations require commanders to verify target classifications and assess proportionality before authorizing attacks, obligations presupposing human decision–making authority at tactically relevant timescales. When autonomous systems execute targeting decisions at speeds exceeding human cognitive capacity for meaningful review, effective human control erodes despite nominal supervisory architecture, creating fundamental incompatibility between international humanitarian law's temporal assumptions and autonomous weapons' operational characteristics.[29]

The three tiers map directly onto three degrees of human control. Tier 1 (human–in–the–loop) systems require operators to directly and actively control the critical functions of selecting and applying force to specific targets, with international humanitarian law's precautionary obligations requiring such human approval before each engagement to ensure compliance with requirements to verify target classifications and assess proportionality.[30] Tier 2 (human–on–the–loop) systems allow autonomous execution of targeting decisions within pre–defined parameters while human supervisors retain the capability to monitor system behavior and intervene to override, veto, or deactivate operations, a supervisory control architecture that requires operators to maintain meaningful human control through real–time oversight despite lacking the ability to approve each individual engagement.[31] Tier 3 (human–out–of–the–loop) systems operate entirely independently after initial deployment without human supervision

29 AP I, arts. 57(2)(a)(i), 57(2)(a)(iii); Vincent Boulanin, Neil Davison, Netta Goussac & Marta Bo, *Autonomous Weapon Systems and International Humanitarian Law: Findings, Recommendations and Remaining Challenges* 3–4 (SIPRI 2021) [hereinafter Boulanin et al., AWS and IHL], https://www.sipri.org/publications/2021/other-publications/autonomous-weapon-systems-and-international-humanitarian-law; Kwik et al., *Controlling Military AI*, *supra* note 26, at 19–20.

30 AP I, art. 57(2)(a)(i); Vincent Boulanin, Neil Davison, Netta Goussac & Moa Peldán Carlsson, Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control, SIPRI Report, at 28 (June 2020) [hereinafter Boulanin et al., *Limits on Autonomy*].

31 Boulanin et al., *Limits on Autonomy*, *supra* note 30, at 28–29; Arthur Holland Michel, *The Black Box, Unlocked: Predictability and Understandability in Military AI*, UNIDIR Report, at 13, 17 (2020) [hereinafter Michel, *The Black Box*].

throughout operational execution, with humans unable to monitor specific decisions or control individual targeting choices.[32]

The pluralistic legal traditions relevant to autonomous weapons governance — including Indigenous, Ubuntu, and Islamic conceptions of intergenerational responsibility — supply normative content for the Human Security by Design framework developed in Chapter 5, Section I. They play no role in the *lex lata* analysis of Chapters 1 through 4 and are not analyzed here. One further dimension of Tier 3 autonomous cyber weapons requires identification before the classification framework is complete: the structure of their supply chain.

A further governance dimension of Tier 3 autonomous cyber weapons — their dependence on commercial artificial intelligence platforms distributed through global supply chains — fragments accountability across developers, cloud infrastructure providers, and deploying States in ways that existing State responsibility frameworks were not designed to reach. Dual-use configurability, which renders an ostensibly defensive machine-learning system into an offensive autonomous weapon through the modification of configuration files alone, compounds the governance deficit by creating regulatory evasion opportunities that neither export controls nor the Article 36 AP I weapons-review framework has been designed to address. The three governance challenges this fragmentation generates — proliferation risk, attribution difficulty, and supply-chain accountability gaps — and the institutional responses they demand are analysed in full in Chapter 5, Sections III and IV.

One doctrinal clarification governs the relationship between the per se prohibition advanced in Chapters 2 and 3 and the regulatory architecture developed in Chapters 5 and 6. The per se prohibition applies to Tier 3 ACW as currently architected. The IAWCO framework does not license

32 Boulanin et al., *Limits on Autonomy*, *supra* note 30, at 28; Michel, *The Black Box*, *supra* note 31, at 13.

the deployment of systems satisfying the Tier 3 definition; it establishes a pre-deployment certification regime whose function is precisely to determine whether a candidate system, through embedded architectural restraints — kill-switches, geofencing, temporal deactivation logic, and human verification loops — has been redesigned to no longer satisfy the Tier 3 definition. A system that passes IAWCO certification has, by definition, been reclassified below the Tier 3 threshold. The regulatory architecture of Chapters 5 and 6 is not an exception to the per se prohibition; it is the institutional mechanism by which that prohibition is operationalized.

Having established the technological architecture and legal classification of Tier 3 autonomous cyber weapons, the following chapters demonstrate what fails when the human decision-maker this architecture eliminates is absent from the moment of engagement — beginning with the *jus contra bellum* frameworks that govern the resort to force and proceeding to the IHL rules that govern its conduct.

Chapter 2
Autonomous Cyber Weapons and the Structural Incapacity of International Law: The Jus Contra Bellum Dimension

This chapter asks a single question: whether the law governing the resort to force — the *jus contra bellum* — can accommodate autonomous cyber weapons whose deployment is authorized at one moment, but whose effects are generated autonomously at another.

This chapter employs the term *jus contra bellum* rather than the traditional *jus ad bellum* to foreground the Charter's primarily prohibitory structure: the prohibition in Article 2(4) is the primary rule; the right of self-defense under Article 51 is a narrow, derogating exception. It proceeds through the prohibition on the use of force under Article 2(4) UN Charter, the self-defense framework of Article 51 UN Charter and the Caroline conditions, the law of countermeasures, and concludes with the armed conflict threshold and the Article 36 AP I weapons-review obligation as the procedural consequence of crossing it, and establishes a single structural finding: each of these frameworks presupposes a State decision-maker who is present when force is applied, and the temporal disaggregation between authorization and autonomous activation defeats that presupposition. This finding — the pre-authorization deficit — is the foundation on which Chapter 3 builds. Where this chapter shows that the resort to force cannot be lawfully decided in advance by a human who is then absent at the moment of force, Chapter 3 shows that the conduct of hostilities under Additional Protocol I fails for the same structural reason. The two chapters are one argument in two registers: the human the law presupposes is, in both, eliminated by design.

The *jus contra bellum* framework — the Charter's rules governing the resort to force — requires a State decision-maker who is present when force is applied. Chapter 1 established what Tier 3 autonomous cyber weapons are. This Chapter demonstrates what they defeat. The Caroline conditions of necessity and imminence presuppose a human decision-maker present at the moment force is applied. The ARSIWA countermeasures regime presupposes that a deploying State can bound its weapon's effects in direction, reversibility, and proportionality. Tier 3 architecture, by disaggregating deployment authorization from autonomous activation by days, months, or years, structurally defeats each presupposition. The result — the pre-authorization problem — is not remediable through better practice. It is an architectural feature of the weapon.

The finding this Chapter establishes is the premise Chapter 3 requires: if the decision to apply force cannot be lawfully taken in advance by a human who is then absent at the moment of execution, the rules governing the *conduct* of that force — distinction, proportionality, precaution — fail for the same structural reason. Chapter 2 proves the defeat of the *jus contra bellum*. Chapter 3 proves the defeat of the *jus in bello*. Together they constitute the dissertation's central doctrinal finding.

The red line of this Chapter: the jus contra bellum presupposes a State decision-maker present at the moment force is applied. Tier 3 systems eliminate that decision-maker before that moment arrives. Every section that follows demonstrates one legal consequence of that elimination.

I. The Jus Contra Bellum Dimension: The Pre-Authorization Problem

The structural defect this dissertation identifies — the elimination of the human decision-maker at the moment of lethal effect — defeats not one body of law but two. This Chapter establishes that defeat for the *jus contra bellum*. Sections I through IV demonstrate that the Charter's foundational requirements — the Article 2(4) prohibition, the Article 51 self-defense right, the Caroline conditions of necessity and imminence, and the ARSIWA countermeasures regime — cannot accommodate the temporal disaggregation that Tier 3 architecture generates by design. Section V identifies the armed conflict threshold at which IHL engages as *lex specialis* and establishes the Article 36 AP I review obligation as the procedural consequence of crossing it. The same structural defect that defeats the *jus contra bellum* defeats the *jus in bello* under AP I — a demonstration carried in full in Chapter 3.

The *jus contra bellum* framework governing autonomous cyber weapons is structured by the binary architecture of the United Nations Charter. Article 2(4) prohibits Member States from threatening or using force against the territorial integrity or political independence of any State,[33] or in any other manner inconsistent with the Purposes of the United Nations, subject to the exception in Article 51 for individual and collective self-defense against an armed attack, with the Caroline conditions of necessity and imminence providing the customary threshold for anticipatory action.[34] This framework presupposes the

33 That architecture presupposes a decision-maker who, at the moment force is initiated, can assess necessity, proportionality, and the existence of an armed attack. Tier 3 deployment destroys that presupposition: the initiating act -deployment authorization-precedes target-selection and attack execution by hours, days, or an indeterminate dormancy period. The central question is therefore whether a deployment authorization issued in advance of verifiable conditions can satisfy the Charter's threshold requirements, and if not, what follows for anticipatory self-defence, collective security authorization, and the pre-authorization of autonomous offensive cyber operations.

34 U.N. Charter art. 2(4), art. 51; Letter from Daniel Webster, U.S. Sec'y of State, to Lord Ashburton (Aug. 6, 1842), *in* 4 Hunter Miller, *Treaties and Other International Acts of the United States of America* 536 (Gov't Printing Office 1934) (necessity of self-defence must be "instant, overwhelming, leaving no choice of means, and no moment for deliberation"), https://avalon.law.yale.edu/19th_century/br-1842d.asp; Michael N. Schmitt et al., *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* r. 71 cmt. 5 (2d ed. 2017).

presence of a decision-maker who, at the moment force is initiated, can assess necessity, proportionality, and the existence of an armed attack. Tier 3 deployment undermines this presupposition, as the initiating act — deployment authorization — precedes target selection and attack execution by hours, days, or an indeterminate period of dormancy. The central issue is whether a deployment authorization issued in advance of verifiable conditions can satisfy the Charter's threshold requirements, and if not, what the implications are for anticipatory self-defense, collective security authorization, and the pre-authorization of autonomous offensive cyber operations.

II. The Prohibition on the Use of Force Under Article 2(4) UN Charter and Its Application to Autonomous Cyber Operations

The drafting history of the Charter and the subsequent practice of States and the ICJ confirm that the prohibition extends beyond military invasion and aerial bombardment to encompass all means and methods by which a State exercises coercive force against another State, including autonomous cyber operations that produce the functional equivalent of kinetic armed attack against critical civilian infrastructure.[35] Two interpretive questions are posed for the Article 2(4) of the UN Charter regime by autonomous cyber weapons: first, what threshold of effects must a cyber operation cross in order to qualify as "force" within the meaning of that provision; and second, how is that threshold assessed where the initiating act — deployment authorization — is temporally disaggregated from the autonomous system's subsequent target-selection and engagement decisions by an indeterminate dormancy period during which the conditions triggering

35 Charter of the United Nations art. 2(4), signed June 26, 1945, 1 U.N.T.S. XVI [hereinafter U.N. Charter]; *Military and Paramilitary Activities in and Against Nicaragua* (Nicar. v. U.S.), Merits, Judgment, 1986 I.C.J. Rep. 14, ¶¶ 190–191 (June 27) [hereinafter Nicaragua] (confirming that the article 2(4) prohibition constitutes a rule of *jus cogens* encompassing a spectrum of coercive acts extending beyond military invasion and distinguishing "the most grave forms of the use of force" from "other less grave forms"); Tallinn Manual 2.0, *supra* note 18, r. 68 cmt. ¶¶ 2, 4 (confirming, on the basis of the Charter's *travaux préparatoires*, a presumption of illegality for any threat or use of force inconsistent with the purposes of the United Nations, and that the prohibition applies irrespective of the means or actor employed)).

autonomous activation remain, by design, unverifiable by the deploying State's legal advisers at the moment of authorization.

1. The Scale–and–Effects Test: Physical Effects Threshold and Divergent State Practice

The ICJ, in *Military and Paramilitary Activities in and Against Nicaragua*, drew a categorical distinction between the most grave forms of the use of force, constituting armed attacks, and "other less grave forms," thereby confirming that Article 2(4) of the Charter encompasses a spectrum of coercive acts whose common characteristic is the capacity to produce effects equivalent in scale and gravity to those caused by conventional military means, rather than any particular technical means of delivery.[36] The *Oil Platforms* judgment reinforced this methodology, the Court confirming that the necessity condition of self-defence is "strict and objective, leaving no room for any 'measure of discretion'," assessed with reference to the specific defensive measures taken against specific targets, the Court finding that the U.S. had not established that attacking the Iranian platforms specifically was necessary to respond to the specific attacks invoked, notwithstanding the existence of a broader pattern of Iranian hostility.[37] The Tallinn Manual 2.0 codified the scale-and-effects methodology in Rules 69–71, with Rule 69 establishing the threshold proposition that cyber operations constitute a use of force when their "scale and effects are comparable to non-cyber operations rising to the level of a use of force"; the commentary to Rule 69 identifies eight non-exhaustive factors bearing on that assessment, severity, immediacy, directness, invasiveness, measurability of effects, military character, State involvement, and presumptive legality, with no single factor being determinative and the assessment proceeding holistically.[38]

---

36 *Nicaragua*, *supra* note 35, at 101, ¶ 191.

37 Oil Platforms (Iran v. U.S.), Merits, Judgment, 2003 I.C.J. Rep. 161, 186, ¶ 73 (Nov. 6).

38 Tallinn Manual 2.0, *supra* note 18, r. 69, at 330, cmt. ¶¶ 9–10, at 334–36; *id.* r. 70, at 337; *id.* r. 71, at 339, cmt. ¶¶ 6–7, at 340–41.

State practice regarding the application of Article 2(4) to cyber operations remains sharply divided, as Table 1 below demonstrates, with no consensus achieved through either the Group of Governmental Experts or the Open-Ended Working Group processes.

2. Critical Infrastructure Attacks as "Force" under Article 2(4) UN Charter: The Qualification Threshold

In light of the unresolved State practice surveyed in Section II.1, a specific and jurisprudentially grounded claim is advanced: autonomous cyber weapons deployed against critical civilian infrastructure, including electrical distribution systems, water treatment facilities, and hospital communications networks, constitute a use of "force" within the meaning of Article 2(4) of the UN Charter. This holds even in the absence of immediate physical destruction, provided that the disruption results in humanitarian consequences functionally equivalent in severity and irreversibility to those caused by a kinetic attack.

The doctrinal basis for this claim rests on three convergent authorities.

First, the Nuclear Weapons Advisory Opinion confirmed the principle of technological neutrality, the Court holding that the laws of armed conflict apply "to all forms of warfare and to all kinds of weapons, those of the past, those of the present and those of the future," thereby rendering the absence of explicit Charter reference to cyber means legally irrelevant.[39]

Second, the Tallinn Manual 2.0 commentary to Rule 69 identifies severity, assessed with reference to consequences impinging on critical national interests, including critical infrastructure, as "the most significant factor" in the scale-and-effects analysis, and the commentary confirms that the nature of the target, such as critical

39 Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 1996 I.C.J. 226, ¶ 86 (July 8) [hereinafter Nuclear Weapons Advisory Opinion] ("the established principles and rules of humanitarian law applicable in armed conflict . . . apply to all forms of warfare and to all kinds of weapons, those of the past, those of the present and those of the future").

infrastructure, bears independently on that assessment, regardless of whether physical infrastructure is physically destroyed.[40]

Third, the documented operational record confirms that cyber operations achieving functional denial of critical infrastructure services produce humanitarian consequences indistinguishable in severity from kinetic attacks on the same infrastructure: the 2015 *BlackEnergy/KillDisk* operation leaving 225,000 civilians without electrical power in sub–zero temperatures,[41] and the 2016 *Industroyer/CRASHOVERRIDE* attack on Ukraine's transmission–level grid,[42] each demonstrate that functional disruption of life–sustaining civilian systems achieves, at the level of humanitarian effect, what kinetic bombardment would have produced, without the physical damage that the traditional threshold analysis has required.

Proceeding on the dual methodological assumption stated in the Introduction, Section IV, that the reported operational facts of Operation Epic Fury are accurate and that the deployed systems satisfied the Tier 3 criteria of Chapter 1, Section I, the temporal disaggregation resulting from autonomous pre-deployment compounds this threshold problem at a structural level. When a State deploys an autonomous system pre-programmed to activate upon detecting specified triggering conditions in a target network, the deployment authorization is separated in both time and causation from the specific acts that may ultimately constitute uses of force. These acts are generated by the system's algorithmic processes in circumstances that no human decision-maker authorized or could contemporaneously assess. The prohibition in Article 2(4) of the UN Charter applies at the moment the use of force

40 Tallinn Manual 2.0, *supra* note 18, r. 69 cmt. ¶¶ 9(a), 10 (identifying severity as "the most significant factor" in the scale-and-effects assessment, with consequences impinging on "critical national interests," including the "nature of the target (such as critical infrastructure)," bearing on that assessment irrespective of whether physical destruction results).

41 Lee, Assante & Conway, *supra* note 2, at iv, 1, 6–8; NCCIC/ICS-CERT Alert IR-ALERT-H-16-043-01AP, *supra* note 2, at 1, 3–5.

42 Cherepanov, *supra* note 4, at 2, 5, 11–12; Dragos, Inc., *supra* note 4, at 4–5, 11, 20–21.

occurs, not at the moment of deployment authorization. However, at the moment of activation, the State that issued the authorization will lack a human decision-maker to evaluate whether the conditions for a lawful use of force exist, to apply the scale-and-effects methodology to the prevailing circumstances, or to abort the operation if the necessity and proportionality thresholds are not satisfied. This structural incapacity, referred to as the "pre-authorization problem," is not a procedural deficiency that can be remedied by improved weapons review protocols; rather, it is an architectural feature of Tier 3 autonomous systems that renders the Charter's use-of-force regime structurally inapplicable as a contemporaneous legal constraint on each specific act of force. The implications for self-defense regimes are examined in Section III of this chapter.

III. Anticipatory Self–Defense, the Caroline Conditions, and the Pre-Authorization Deficit under Article 51 UN Charter

The distinction between automated and autonomous systems established in Chapter 1 is legally determinative of the pre-authorization problem examined in this Section.[43] Tier 3 architecture eliminates the human decision–maker at the operative moment of autonomous activation by design, and with it all three conditions of the Caroline formula: necessity cannot be assessed at the moment force is applied, imminence cannot be evaluated at the moment the trigger fires, and proportionality cannot be calibrated against circumstances that no human decision-maker is present to assess at the operative moment of activation.[44]

The Caroline doctrine establishes two cumulative conditions for lawful anticipatory self-defense: the necessity of the defensive act must be

43 *See supra* Chapter 1, Section I.1 and accompanying text.

44 *See supra* Chapter 1, Sections I.2, II.1; Letter from Daniel Webster, U.S. Sec'y of State, to Lord Ashburton (Aug. 6, 1842), *in* 4 Hunter Miller, *Treaties and Other International Acts of the United States of America* 536 (Gov't Printing Office 1934) ("necessity of self-defence must be 'instant, overwhelming, leaving no choice of means, and no moment for deliberation'") [hereinafter Webster-Ashburton Correspondence].

"instant, overwhelming, leaving no choice of means, and no moment for deliberation," (hereinafter the "Caroline formula"), and the response must not be excessive in relation to the threat.[45] The ICJ has affirmed the continued legal force of both conditions: as the Court established in *Military and Paramilitary Activities in and Against Nicaragua (Nicaragua v. United States of America)*, Merits, 1986 ICJ Rep. 14, para. 194, and applied in *Oil Platforms (Iran v. United States)*, Merits, 2003 ICJ Rep. 161, para. 43, the necessity requirement "is strict and objective, leaving no room for any measure of discretion," while in the same judgment it endorsed the established customary rule that "whether the response to the armed attack is lawful depends on observance of the criteria of the necessity and the proportionality of the measures taken in self-defense."[46] Both limbs presuppose a human decision–maker capable of evaluating, at the moment of contemplated force, whether the threshold conditions are satisfied in the specific circumstances then obtaining, not as modeled at the moment of system programming, but as they exist at the moment of activation.

The right of self-defense under Article 51 of the UN Charter, even as applied to non-State armed attacks, is subject to a territorial limitation that generates an additional and independent structural obstacle for autonomous cyber weapons: the forcible response must be justified separately vis-à-vis every State on whose territory force is exercised. As Kreß observes, "from a legal perspective... such an automatic extension of the geographical scope of the right of self-defense is unwarranted

45 Letter from Daniel Webster to Lord Ashburton (Aug. 6, 1842), *in* 4 Hunter Miller ed., Treaties and Other International Acts of the United States of America 536–37 (Gov't Printing Office 1934) ("the necessity of that self-defense is instant, overwhelming, and leaving no choice of means, and no moment for deliberation"); *accord* Letter from Lord Ashburton to Daniel Webster (July 28, 1842), *in id.* ("that necessity of self-defense, instant, overwhelming, leaving no choice of means"); Letter from Daniel Webster to Henry Fox (Apr. 24, 1841), *in id.* ("a necessity of self-defense, instant, overwhelming, leaving no choice of means, and no moment for deliberation") [collectively cited *supra* note 44 as Webster-Ashburton Correspondence].

46 Oil Platforms (Islamic Republic of Iran v. United States of America), Merits, Judgment, 2003 I.C.J. 161, 186 ¶ 73 (Nov. 6) ("strict and objective, leaving no room for any 'measure of discretion'"); *id.* ¶ 74 (quoting *Military and Paramilitary Activities in and Against Nicaragua* (Nicar. v. U.S.), Merits, Judgment, 1986 I.C.J. 14, 103 ¶ 194 (June 27)); *see also Nicaragua*, *supra* note 35, at 94 ¶ 176 ("self-defense would warrant only measures which are proportional to the armed attack and necessary to respond to it . . . a rule well established in customary international law").

because it loses sight of the necessary inter-State reading of the right to self-defense under current international law"; accordingly, "it must always be shown that the non-State armed violence emanating from the State on whose territory armed force is used has reached (or is about to reach) the required quantitative threshold."[47] A Tier 3 autonomous cyber weapon pre-authorized to propagate across interconnected networks cannot discharge this obligation: by the moment of autonomous territorial expansion, the four cumulative conditions governing transnational self-defense, intensity, subsidiarity, proportionality, and territorial limitation, cannot be satisfied jurisdiction-by-jurisdiction at the moment of each autonomous territorial entry.

Applied to autonomous cyber operations, the necessity limb is structurally defeated at two independent levels. By the moment of activation, the necessity evaluation performed at deployment is temporally and factually stale: it cannot satisfy the necessity standard formulated by Secretary of State Webster in his 1842 correspondence with Lord Ashburton, requiring that necessity be "instant, overwhelming, leaving no choice of means, and no moment for deliberation", because the decision was not made at the moment force was applied but at a prior, materially different moment. Second, the imminence standard, requiring that an armed attack be about to occur at the moment of the responsive act, is structurally unavailable for a system that activates on pre–programmed network–load triggers, with no

47 Claus Kreß, *Transnational Armed Conflicts*, 15 J. Conflict & Sec. L. 245, 250–52 (2010) (four cumulative conditions at 249–52; territorial limitation at 250–51: "the right to self-defense against a non-State armed attack justifies a forcible response only on the territory of the State from whose territory the non-State armed attack occurs"; "the exercise of the right of self-defense must be specifically justified vis-à-vis every State on whose territory self-defense action is taken"); UN Charter, *supra* note 35, art. 51; Legal Consequences of the Construction of a Wall in the Occupied Palestinian Territory, Advisory Opinion, 2004 I.C.J. Rep. 136, ¶ 139 (July 9); Armed Activities on the Territory of the Congo (Dem. Rep. Congo v. Uganda), Merits, Judgment, 2005 I.C.J. Rep. 168, 223, ¶¶ 146–147 (Dec. 19) (both cited in Kreß, *supra*, at 248 n.10).

human present to assess whether those conditions indicate an imminent armed attack at the operative moment of activation.[48]

The Tallinn Manual 2.0 provides indirect doctrinal support for this conclusion: the International Group of Experts concluded that the introduction of malware activating upon pre-programmed triggers constitutes an "attack" under the law of armed conflict, thereby triggering the full panoply of precautionary obligations under Article 57–58 of AP I, irrespective of whether the system has activated at the moment of assessment, provided the intended consequences meet the requisite harm threshold. That classification does not itself resolve the pre-authorization question, which is this dissertation's contribution; it does, however, confirm that the moment of introduction, not activation, is legally operative, which reinforces the conclusion that a deployment authorization temporally disaggregated from activation cannot discharge the Caroline conditions.[49] A Tier 3 autonomous cyber weapon cannot satisfy the Caroline formula, not because of an evidentiary deficit, but because its architecture structurally eliminates the human decision–maker whom both limbs of the formula presuppose.[50]

The objection most likely to be pressed by States defending Tier 3 pre-deployment takes the following form: the Caroline formula was developed for State-to-State self-defense in a pre-digital era; its "instant, overwhelming, leaving no choice of means, and no moment for deliberation" standard was designed to exclude leisurely preventive war,

48 Webster-Ashburton Correspondence, *supra* note 44 (necessity must be "instant, overwhelming, leaving no choice of means, and no moment for deliberation"); UN Charter, *supra* note 35, art. 51; Nuclear Weapons Advisory Opinion, *supra* note 39, ¶ 41 (confirming continued legal force of the Caroline conditions); Daniel Bethlehem, *Self-Defense Against an Imminent or Actual Armed Attack by Non-State Actors*, 106 Am. J. Int'l L. 769, 775–76 (2012) [hereinafter Bethlehem, *Self-Defense*].

49 Tallinn Manual 2.0, *supra* note 18, r. 92 cmt. ¶ 16, at 418 ("the introduction of malware or production-level defects that are either time-delayed or activate on the occurrence of a particular event is an attack when the intended consequences meet the requisite threshold of harm . . . irrespective of whether they are activated"); *see also id.* r. 92 cmt. ¶ 15, at 418 (cyber operation causing foreseeable collateral damage at the requisite level constitutes an attack "to which the relevant law of armed conflict applies, particularly that regarding proportionality (Rule 113)"); *cf. id.* r. 113 cmt. ¶¶ 3, 6–7, at 463–64 (indirect and cascading effects to be factored into the proportionality calculation; uncertainty as to collateral damage does not reduce the calculation threshold).

50 Oil Platforms (Iran v. U.S.), Merits, Judgment, 2003 I.C.J. Rep. 161, 186, ¶¶ 73–74 (Nov. 6); AP I, art. 51(5)(b); *cf.* AP I, art. 52(2); Boulanin et al., *Limits on Autonomy*, *supra* note 30, at 7–8.

not to require human presence at each autonomous activation in a machine-speed threat environment. On this reading, a State that deliberates carefully at the deployment authorization stage, conducting full necessity and imminence assessments, war-gaming activation scenarios, establishing clear trigger criteria, has satisfied the Caroline conditions at the moment when deliberation was possible, and the subsequent autonomous activation is merely the execution of a lawful decision already made. This argument has intuitive appeal and represents the most plausible legal defense a State would actually advance. It fails for the following reason.

The Caroline formula's "instant" requirement is not merely a prohibition on leisurely preventive war; it is an affirmative requirement that the necessity be assessed at the moment force is applied, not at a prior moment when assessment was convenient. Webster's formulation, "instant, overwhelming, leaving no choice of means, and no moment for deliberation", describes the conditions that must obtain at the moment of the defensive act, not at a prior planning stage. The ICJ confirmed this temporal specificity in Oil Platforms when it required that the specific defensive measures taken against specific targets be necessary to respond to the specific attacks invoked, a standard assessed at the moment of the response, not at the moment of its authorization. A deployment authorization that satisfied the Caroline conditions at the time it was issued cannot satisfy them at the moment of autonomous activation if the conditions have changed materially in the interval, and temporal disaggregation, by design, creates precisely that interval. The argument that pre-deployment deliberation satisfies the Caroline formula would effectively permit States to conduct leisurely advance planning and then execute it through autonomous systems under the label of "instant" self-defense, inverting the formula's protective logic.

IV. Countermeasures, Forcible Autonomous Cyber Operations, and the Structural Tension

Conventional kinetic autonomous weapon systems operate within bounded physical spaces shaped in practice by factors such as fuel capacity, terrain navigability, and communications range, which together constrain their effective engagement envelopes. U.S. Department of Defense policy requires that autonomous and semi–autonomous weapon systems complete engagements only within defined timeframes, geographic areas, and other operational constraints consistent with commander and operator intentions, and mandates that systems terminate the engagement or obtain additional operator input before continuing where those boundaries cannot be maintained.[51] An autonomous aerial drone cannot spontaneously expand its operational domain to adjacent regions absent human redeployment. Physical constraints contain the effects of kinetic systems within predictable geographic boundaries, enabling commanders to assess proportionality based on spatially bounded civilian populations under established principles that prohibit attacks causing excessive incidental civilian harm relative to the anticipated concrete and direct military advantage. In practice, targeting staff can therefore model the likely footprint of kinetic effects with a relatively high degree of confidence, drawing on known weapon characteristics, terrain data, and intelligence on civilian presence.[52]

The structural contrast with autonomous cyber weapons is immediate and analytically decisive. A self-propagating autonomous cyber weapon operates in a heterogeneous, borderless network environment in which propagation behavior cannot be spatially bounded at the moment of deployment authorization, with the consequence that each of the three

51 AP I, art. 36; DoD Directive 3000.09, *supra* note 7, § 4.

52 AP I, arts. 51(5)(b), 57(2)(a)(i)–(iii); ICRC IHL Challenges Report 2024, *supra* note 19, at 7.

cumulative ARSIWA countermeasures conditions encounters a structural, rather than merely evidentiary, barrier.

The directionality condition, that countermeasures be directed against the responsible State, is structurally imperiled by the propagation dynamics established in Chapter 1, Sub-section I.2(a).[53] A self-propagating autonomous cyber weapon deployed against the infrastructure of the responsible State will, by design, traverse shared network dependencies connecting that State's infrastructure to civilian systems, third-party platforms, and internationally operated critical services. The injured State cannot guarantee, at the moment of deployment authorization, that the weapon's effects will remain territorially and functionally bounded to the responsible State, because emergent propagation behavior in heterogeneous network environments cannot be excluded *ex-ante*. A countermeasure that foreseeably extends effects to third States or to civilian infrastructure in the responsible State beyond the scope of the injury suffered violates the directionality condition and constitutes an independent internationally wrongful act.[54]

The reversibility condition imposes an independent and absolute structural barrier. Article 49(3) ARSIWA requires that countermeasures "shall, as far as possible, be taken in such a way as to permit the resumption of performance of the obligations in question." The ARSIWA Commentary confirms that this condition reflects the general principle that States should prefer reversible measures when alternatives exist.[55] Autonomous cyber weapons exhibiting irreversibility through self-propagating code traversing third-party infrastructure beyond the deploying State's technical control cannot satisfy this condition once propagation has exceeded the State's reach. Even where technical termination mechanisms, "kill switches" or remote update functions, are

53 ARSIWA, *supra* note 21, art. 49(1) & cmt. ¶ 3; *see supra* Chapter 1, Section IV.1.

54 ARSIWA, *supra* note 21, art. 49 cmt. ¶ 4.

55 ARSIWA, *supra* note 21, art. 49(3) & cmt. ¶ 9.

embedded at the design stage, their effectiveness depends on continued network connectivity, intermediary cooperation, and the absence of adversarial countermeasures, conditions that cannot be guaranteed in an active conflict environment. Where the deploying State loses practical control over the weapon's propagation, the ARSIWA reversibility requirement is not merely difficult to satisfy; it is structurally unavailable from the moment technical control is lost.[56]

The proportionality condition, that countermeasures must be commensurate with the injury suffered, taking into account the gravity of the internationally wrongful act and the rights in question, as codified in Article 51 ARSIWA, presents a further structural obstacle for Tier 3 autonomous cyber countermeasures. Proportionality under ARSIWA requires contemporaneous calibration of the countermeasure's effects against the specific injury; this calibration is structurally impossible where the weapon's propagation behavior cannot be bounded at the moment of deployment authorization, because the total quantum of harm produced by an autonomous self-propagating system remains unknown, and unknowable, to the deploying State at the moment the calibration must be performed.

The prohibition on forcible countermeasures is categorical and admits of no exception. Article 50(1)(a) ARSIWA expressly provides that countermeasures shall not affect "the obligation to refrain from the threat or use of force as embodied in the Charter of the United Nations." The ARSIWA Commentary confirms that this prohibition reflects a peremptory norm and is "widely supported by extensive literature and consistent judicial decisions," drawing on the Declaration on Principles of International Law concerning Friendly Relations and Cooperation among States, which proclaimed that "States have a duty to refrain from

56 Hui Chen et al., *Countermeasures in Cyberspace* 15–16 (GCSP Nov. 2024), https://www.gcsp.ch/publications/countermeasures-cyberspace (last visited Apr. 16, 2026).

acts of reprisal involving the use of force."[57] Where an autonomous cyber countermeasure produces effects qualifying as a "use of force" under Article 2(4) of the UN Charter, the ARSIWA countermeasures regime is categorically unavailable, irrespective of the injury's gravity or the deploying State's subjective intent.[58]

Having demonstrated that the jus contra bellum's Caroline and Article 51 framework cannot accommodate the pre-authorization problem, this chapter now turns to its core analysis: whether the same structural defect defeats the precautionary architecture of AP I, beginning with the threshold question of when that architecture applies.

V. The Armed Conflict Threshold, Temporal Disaggregation, and the Article 36 AP I Review Obligation

The *jus contra bellum* analysis of Sections I through IV has established that Tier 3 autonomous cyber weapons defeat the Charter's use-of-force framework at the moment of authorization. A distinct but complementary question now arises: when does the deployment or activation of such a weapon cross the threshold that triggers IHL as *lex specialis*, and what obligations attach at that moment? The answer determines whether the full precautionary architecture of AP I — the rules governing distinction, proportionality, and precaution — engages at all. Section V addresses the threshold question in three steps: the general armed conflict threshold and its application to cyber operations; the Article 36 AP I weapons-review obligation as the procedural consequence of crossing it; and the temporal disaggregation problem that distinguishes autonomous cyber operations from all prior means of warfare to which those rules have been applied.

57 ARSIWA, *supra* note 21, art. 50(1)(a) & cmt. ¶ 5; G.A. Res. 2625 (XXV), annex, princ. 1, ¶ 9, U.N. Doc. A/RES/2625(XXV) (Oct. 24, 1970), https://undocs.org/A/RES/2625(XXV.

58 *Nicaragua*, *supra* note 35, ¶ 249; *see supra* Chapter 1, Section II.1.

1. The Armed Conflict Threshold and Technological Neutrality

The threshold question concerns when cyber operations trigger the application of IHL as *lex specialis*, displacing peacetime legal regimes. The ICJ confirmed in its *Nuclear Weapons* Advisory Opinion that IHL applies as a specialized law of armed conflict, establishing its operational primacy once the armed conflict threshold is crossed.[59] The Court further elaborated the lex specialis principle in its Wall Advisory Opinion, confirming that IHL and international human rights law apply concurrently to situations of armed conflict, with IHL constituting the more specific body of law governing the conduct of hostilities.[60]

The determination of when cyber operations occur "in armed conflict" requires applying the framework established by the International Criminal Tribunal for the former Yugoslavia in *Prosecutor v. Tadić*, which articulated the authoritative test: "armed conflict exists whenever there is a resort to armed force between States or protracted armed violence between governmental authorities and organized armed groups or between such groups within a State."[61] This definition encompasses both international armed conflicts between States under Common Article 2 of the Geneva Conventions and non–international armed conflicts involving non–State actors that meet the organizational and intensity thresholds under Common Article 3 of the Geneva Conventions.[62]

---

59 Nuclear Weapons Advisory Opinion, *supra* note 39, ¶ 25.

60 Legal Consequences of the Construction of a Wall in the Occupied Palestinian Territory, Advisory Opinion, 2004 I.C.J. Rep. 136, ¶ 106 (July 9) ("As regards the relationship between international humanitarian law and human rights law, there are thus three possible situations: some rights may be exclusively matters of international humanitarian law; others may be exclusively matters of human rights law; yet others may be matters of both these branches of international law. In order to answer the question put to it, the Court will have to take into consideration both these branches of international law, namely human rights law and, as *lex specialis*, international humanitarian law."), https://www.icj-cij.org/case/131 (last visited Apr. 16, 2026).

61 Prosecutor v. Tadić, Case No. IT-94-1-AR72, Appeals Chamber, Decision on the Defense Motion for Interlocutory Appeal on Jurisdiction, ¶ 70 (Int'l Crim. Trib. for the Former Yugoslavia Oct. 2, 1995) ("an armed conflict exists whenever there is a resort to armed force between States or protracted armed violence between governmental authorities and organized armed groups or between such groups within a State"), https://www.icty.org/x/cases/tadic/acdec/en/51002.htm (last visited Apr. 16, 2026) [hereinafter Tadić Jurisdiction Decision].

62 GC IV, *supra* note 68, arts. 2–3; Tadić Jurisdiction Decision, *supra* note 61, ¶ 70.

The ICRC emphasizes that IHL's applicability depends not on the technological means employed but on the existence of an armed conflict that meets these criteria, confirming that the legal framework applies to "all means and methods of warfare used by parties to an armed conflict, regardless of the technology involved."[63]

Two clarifications sharpen the analysis. First, the armed conflict threshold does not require that hostilities be conducted exclusively through kinetic means; cyber operations may occur within otherwise conventional armed conflicts, and cyber means may independently satisfy the *Tadić* intensity requirements where effects reach thresholds comparable to those of conventional armed forces. While Tallinn Manual 2.0, a non–binding expert restatement produced by an International Group of Experts convened at the invitation of the NATO Cooperative Cyber Defence Centre of Excellence, without representing the official position of the Centre, NATO, or any participating State, concludes that cyber operations meeting the IHL definition of "attack" constitute engagement in hostilities and may thereby trigger international armed conflict status, this position remains a scholarly interpretation rather than evidence of binding State practice. This doctrinal framework remains largely theoretical because no State has thus far formally invoked armed conflict status solely on the basis of cyber hostilities directed against it. This empirical gap reflects the intensity threshold for non–international armed conflict under the *Tadić*

63 ICRC Cyber Operations Report, *supra* note 23, at 4; ICRC IHL Challenges Report 2024, *supra* note 19, at 59.

test, a threshold that cyber operations conducted independently of kinetic force are unlikely, though not categorically unable, to satisfy.[64]

Second, while the technological modality of harm does not determine IHL applicability, an operation's nexus to armed conflict presents interpretive challenges when cyber operations exhibit temporal disaggregation, initial intrusion separated from the generation of effects, and institutional fragmentation across distinct State organs. Crossing the armed conflict threshold triggers binding obligations at two levels. Substantively, it activates the core targeting restrictions: the prohibition on directing operations against civilians and civilian objects under Article 48 of AP I, the proportionality requirement under Article 51(5)(b) AP I prohibiting attacks expected to cause civilian harm that are excessive in relation to the anticipated military advantage, and the precautionary obligations under Article 57–58 AP I requiring constant care to spare civilians. Procedurally, it activates the Article 36 AP I weapons review obligation, under which High Contracting Parties must determine "whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law."[65] For autonomous cyber weapons, this procedural obligation becomes especially salient because temporal disaggregation may obscure whether the moment of decision for Article 57 AP I purposes occurs at pre–positioning (potentially a peacetime intelligence operation) or at activation (a wartime attack), a doctrinal tension

64 Tallinn Manual 2.0, *supra* note 18, Introduction, at 2 ("Tallinn Manual 2.0 is not an official document, but rather the product of two separate endeavours undertaken by groups of independent experts acting solely in their personal capacity. The Manual does not represent the views of the NATO CCD COE, its sponsoring nations, or NATO. . . . Ultimately, Tallinn Manual 2.0 must be understood only as an expression of the opinions of the two International Groups of Experts as to the state of the law"); *id.* r. 82 cmt. ¶ 11 ("The constituent hostilities may include any combination of kinetic and cyber operations, or cyber operations alone"); *id.* r. 82 cmt. ¶ 16 ("To date, no international armed conflict has been publicly characterised as having been solely precipitated in cyberspace. Nevertheless, the International Group of Experts agreed that cyber operations alone have the potential for crossing the threshold of international armed conflict"); *id.* r. 80 cmt. ¶ 1 ("the law of armed conflict applies to such activities during both international and non-international armed conflicts"). On the distinction between expert restatements and binding State practice in the formation of customary international law, *see Nicaragua*, *supra* note 35, ¶ 183; *Cf.* Tadić Jurisdiction Decision, *supra* note 61, ¶ 70 (intensity threshold for non-international armed conflict and its application to cyber operations.

65 AP I, arts. 36, 48, 51(5)(b), 57–58.

developed in Section V.2 of this Chapter, thereby rendering ex post compliance verification structurally impossible and shifting the protective burden to ex ante Article 36 AP I review.[66]

2. Temporal Disaggregation and the Fragmentation of Article 57's Decision Architecture

Cyber operations introduce a distinct temporal disaggregation between the establishment of access and the generation of effects. Unlike kinetic engagement, where maneuver and fire often occur in close sequence, cyber capabilities often function as "advanced force operations" that precede the main effort by extended intervals. This preparatory phase includes reconnaissance (e.g., mapping a network), seizure of supporting positions (e.g., securing access to key nodes), and the pre–emplacement of capabilities (e.g., implanting malicious code).[67]

The three-component architecture of autonomous cyber weapons — sensors, algorithmic decision modules, and effectors — generates this temporal disaggregation structurally. Algorithmic decision-making processes in autonomous cyber agents employ machine-learning techniques, including online learning from observed outcomes and game-theoretic search over possible action sequences, to determine an optimal course of action under uncertainty, adapting their strategies in response to adversary behavior and environmental conditions without requiring human intervention at each decision point.[68] Contemporary implementations of autonomous systems primarily employ supervised learning, in which deep neural networks are trained on large datasets of labeled examples to learn input-output mappings — for instance, training a network on labeled images until it can classify new images by

66 AP I, art. 57(2)(a).

67 U.S. Dep't of Defense, *Law of War Manual* §§ 16.1.2.1, at 1025 (June 2015, updated July 2023) [hereinafter DoD Law of War Manual]; ICRC IHL Challenges Report 2024, *supra* note 19, at 7.

68 Alexander Kott et al., *Toward Intelligent Autonomous Agents for Cyber Defense: Report of the 2017 Workshop by the NATO Research Group IST-152/RTG*, at 7–9 (U.S. Army Research Lab. Special Report ARL-SR-0395, Apr. 2018), https://arxiv.org/pdf/1804.07646.pdf (last visited Apr. 16, 2026).

computing output scores and adjusting internal parameters to minimize classification error.[69] AWS select and apply force to targets without human intervention, meaning that after initial activation by a human, the application of force is triggered in response to information from the environment received through sensors and on the basis of a generalized target profile, with the user not choosing, or even knowing, the specific person or object that will trigger the system to strike, or precisely when and/or where the strike will occur.[70]

This temporal separation distinguishes autonomous weapon systems from remotely piloted systems, in which human operators maintain continuous communication links, enabling real-time oversight and intervention. The effector component executes force without any real-time human decision at the moment of activation, on the basis of conditions that no Article 57 AP I legal assessment can contemporaneously evaluate. Neither the scale-and-effects threshold nor the armed-attack determination can be applied at the moment force is used; both assessments were performed, if at all, at the earlier moment of deployment authorization.

Spatially, these operations traverse globally distributed infrastructure, often routing through neutral or third–State systems. The ICRC has emphasized that the "interconnectedness" of military and civilian networks means that targeting choices can generate foreseeable reverberating effects beyond the immediate military objective.[71] This spatial propagation mirrors the compliance challenge of autonomous weapon systems (AWS), where the user may not know the specific target or the location of a strike after activation, constraining the ability to assess proportionality in real time.[72]

69 Yann LeCun et al., *Deep Learning*, 521 Nature 436, 436–37 (2015).

70 ICRC IHL Challenges Report 2024, *supra* note 19, at 52.

71 ICRC IHL Challenges Report 2024, *supra* note 19, at 7.

72 *Id.* at 52–53.

Through cyber means, belligerents can infiltrate systems to collect, exfiltrate, modify, encrypt, or destroy data, and can trigger, alter, or manipulate processes controlled by compromised computer systems.[73] These capabilities enable temporal separation between initial intrusion, which may occur months or years before operational use, and the generation of effects, creating distinct phases that maintain covert presence without causing externally observable harm until activation.

3. Temporal Misalignment and the Article 36 AP I Review Obligation

Article 57 of AP I requires those who plan or decide upon an attack to verify that objectives are military targets and to assess expected civilian harm against anticipated military advantage, with these precautionary measures to be taken "at the moment when the decision to attack is made."[74]

Article 57(2)(b) AP I further requires commanders to cancel or suspend attacks "if it becomes apparent" that the target is civilian or that harm would be excessive.[75] This precautionary framework presupposes temporal proximity among the planning, decision, and execution phases because it relies on contemporaneous intelligence and operational capacity to intervene. Traditional military targeting typically exhibits such proximity: commanders assess feasibility and proportionality at the time of the decision based on circumstances then prevailing, and pre–planned targeting cycles allow time for verification procedures. The ICRC Commentary confirms that verification obligations are assessed based on "all circumstances ruling at the time," including available intelligence, collateral damage estimation methodologies, and the

73 ICRC Cyber Operations Report, *supra* note 23, at 3.

74 AP I, arts. 57(2)(a)(i), 57(2)(b); 1987 ICRC AP Commentary, *supra* note 18, ¶ 2198 (interpreting "everything feasible" as "everything that was practicable or practically possible, taking into account all the circumstances at the time of the attack").

75 AP I, art. 57(2)(b); Jean-Marie Henckaerts & Louise Doswald-Beck, 1 CUSTOMARY INTERNATIONAL HUMANITARIAN LAW: RULES r. 19, at 58–62 (Int'l Comm. of the Red Cross & Cambridge Univ. Press 2005) [hereinafter ICRC Customary IHL Study].

anticipated military advantage of the specific operation.[76] Cyber operations, by contrast, exhibit temporal disaggregation: network intrusion, persistent access, and effects generation may be separated by months or years.[77]

This temporal disaggregation generates a doctrinally acute problem. The precautionary calculus required by Article 57 and 58 AP I presumes that decision–makers possess sufficient information to evaluate civilian risk and select means and methods that minimize harm at the time of attack.

Taking the publicly reported facts of Operation Epic Fury as an operational illustration (the Preface explains the epistemic status of this assumption):

Autonomous cyber operations activating months after initial network penetration cannot account for the real-time evolution of humanitarian conditions in the target environment, a dynamic that Article 57 AP I requires decision-makers to assess at the moment of attack, having regard to "all the circumstances at the time of the attack," a standard the Diplomatic Conference deliberately set above the prior threshold of mere reasonable steps.[78] Yet, where access is obtained years earlier,

76 1987 ICRC AP Commentary, *supra* note 18, ¶ 2198 (interpreting "everything feasible" as requiring assessment "taking into account all the circumstances at the time of the attack"), ¶ 2195 (requiring commanders to call for additional information "in case of doubt" based on current intelligence), ¶ 2223 (requiring commanders to "gather all the information available" about civilian presence before attack).

77 *See* FireEye Mandiant, Highly Evasive Attacker Leverages SolarWinds Supply Chain to Compromise Multiple Global Victims With SUNBURST Backdoor (Dec. 13, 2020), https://www.mandiant.com/resources/evasive-attacker-leverages-solarwinds-supply-chain-compromises-with-sunburst-backdoor [hereinafter FireEye Mandiant, SolarWinds] (intrusion began September 2019, malicious code injected February 2020, discovered December 2020 - demonstrating 15-month operational window); U.S. Gov't Accountability Office, GAO-21-288, SolarWinds Cyberattack: Actions Needed for Better Federal Cybersecurity Response and Recovery 1 (2021) [hereinafter GAO-21-288] (confirming December 2020 discovery with "widespread impact on government agencies, critical infrastructures, and the private sector"); Nicolas Falliere, Liam O'Murchu & Eric Chien, W32.Stuxnet Dossier 2, 36 (Symantec Security Response, ver. 1.4, Feb. 2011) [hereinafter Falliere, Murchu & Chien] (multi-phase deployment with temporal separation between initial infection and payload activation).

78 1987 ICRC AP Commentary, *supra* note 18, ¶ 2198 ("Everything feasible" means everything that was practicable or practically possible, taking into account all the circumstances at the time of the attack, including those relevant to the success of military operations); *see also* Draft Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts art. 50 (ICRC 1973), *reprinted in* Official Records of the Diplomatic Conference on the Reaffirmation and Development of International Humanitarian Law Applicable in Armed Conflicts, Geneva, 1974–1977, vol. I (Swiss Fed. Dep't of Foreign Affairs 1978) (ICRC draft juxtaposing "all reasonable steps" (Proposal II) against the subsequently adopted "everything feasible" formulation, demonstrating the Diplomatic Conference's deliberate elevation of the threshold to an objective, operationally concrete standard).

often by different State organs, under different legal mandates, and against a moving target environment, the information needed for "everything feasible" verification may be unavailable at activation, and "the circumstances at the time of the attack" may have shifted in ways that the original intrusion planners could not reasonably anticipate. If a military commander activates a pre–positioned cyber implant during armed conflict, must the commander re–verify target status and re–assess proportionality based on current circumstances, or does the peacetime intrusion satisfy Article 57's obligations? The text of AP I does not explicitly address this scenario, but the precautionary regime's protective logic suggests that verification obligations attach to both phases: initial deployment and final activation.[79]

Attribution concerns sharpen the point at the institutional level. Classical State responsibility doctrine presupposes identifiable conduct attributable to the State through its organs and empowered entities, and it ordinarily operates by reconstructing the act and the chain of control ex post.[80] Temporal separation fragments responsibility across actors who may be subject to distinct domestic and international constraints at different phases, thereby creating a particular type of regulatory gap: not the absence of applicable law, but the inability of IHL's ex-ante precautions to function as designed when preparation and execution are institutionally and temporally decoupled.

The due diligence obligation affirmed in *Corfu Channel*, extended by analogy to autonomous cyber weapons whose propagation

79 AP I, arts. 57(2)(a)(i), 57(2)(b) (requiring "those who plan or decide upon an attack" to "do everything feasible to verify" targets and to "cancel or suspend" attacks if objectives are civilian or harm would be excessive), 58(c) (requiring parties to "take the other necessary precautions to protect the civilian population"); 1987 ICRC AP Commentary, *supra* note 18, ¶ 2198 (interpreting "every-thing feasible" as "that which is practicable or practically possible, taking into account all the circumstances at the time of the attack"), ¶ 2195 (requiring commanders to "call for additional information" "in case of doubt, even if there is only slight doubt"), ¶ 2191 (confirming article 57 imposes "important duty on belligerents" to "always distinguish" between civilian and military objectives); ICRC Customary IHL Study, *supra* note 18, r. 16, at 51 (stating "each party to the conflict must do everything feasible to verify that targets are military objectives").

80 ARSIWA, *supra* note 21, arts. 2, 4, 8 (establishing attribution framework); *Nicaragua*, *supra* note 35, ¶ 115 ("directed or enforced" test); Tadić Appeals Judgment, *supra* note 61, ¶¶ 120, 131 ("overall control" test).

characteristics create foreseeable transboundary risks at the moment of authorization, reinforces this preventive orientation: where ex post verification is structurally impossible, the compliance burden migrates upstream to the Article 36 review stage. The full doctrinal analysis of this due diligence obligation and its extension to autonomous deployment is developed in Chapter 5, Section I.4.

The same structural defect — the elimination of the human decision-maker before force arrives — that defeats the *jus contra bellum* is shown, in Chapter 3, to defeat the *jus in bello* under Additional Protocol I by the same logic and for the same reason.

Chapter 3
Autonomous Cyber Weapons and the Structural Incapacity of International Law: The Jus In Bello Dimension

Chapter 2 established that the *jus contra bellum* cannot accommodate Tier 3 autonomous cyber weapons: the Article 2(4) prohibition, the Article 51 self-defense right, the Caroline conditions, and the ARSIWA countermeasures regime each presuppose a State decision-maker present at the moment force is applied, and Tier 3 architecture eliminates that decision-maker in advance.

This Chapter asks the next question: once the attack threshold is crossed, can the rules governing the conduct of hostilities — distinction, proportionality, and precaution under Additional Protocol I — be lawfully discharged by a Tier 3 autonomous cyber weapon?

The answer is no, and for the same structural reason. Each of those rules also presupposes a human being present at the operative moment of lethal effect — a human capable of verifying a target, weighing incidental harm, and cancelling an attack. Tier 3 systems eliminate that human by design.

The analysis proceeds through three movements.

The first establishes the jurisdictional threshold: whether autonomous cyber operations constitute "attacks" under Article 49(1) AP I, engaging the full protective architecture of Articles 48 through 58.

The second demonstrates, across distinction, proportionality, and precaution in turn, that each of those core targeting obligations is structurally incapable of operational implementation by a Tier 3 system — and establishes the Algorithmic Transparency Obligation, an original contribution of this dissertation, as what existing law already requires.

The third demonstrates that the same structural incapacity extends to the special protections for medical units, objects indispensable to civilian survival, and works containing dangerous forces.

The red line of this Chapter: the targeting regime of AP I presupposes a human who can verify, weigh, and cancel at the moment force is applied. Tier 3 systems eliminate that human architecturally.

The accountability consequences of this elimination are analyzed in Chapter 4.

The design architecture that restores what the law presupposes is developed in Chapter 5.

I. The Violence Threshold under Article 49(1) AP I: Functional Interpretation, State Practice, and the Stuxnet Precedent

The operationalization of IHL targeting rules turns on the definition of "attack" in Article 49(1) AP I as "acts of violence against the adversary, whether in offense or in defense."[81] This definition is not merely taxonomic; it is jurisdictional in the conduct of hostilities because it serves as the doctrinal gateway to the rules on distinction, proportionality, and precautions in attack. In the cyber context and even more so where capabilities are autonomous or self–propagating, this gateway function assumes decisive importance: if an operation is characterized as an attack, the full protective machinery of Article 48–58 AP I is triggered; if not, the operation may still be regulated by other bodies of international law (sovereignty, *jus contra bellum*, human rights), but it will not be disciplined by the core targeting regime.[82]

81 AP I, art. 49(1).

82 AP I, arts. 48–58 (distinction, proportionality, precautions).

1. The Violence Threshold: Functional Interpretation under Article 31(1) VCLT

The text of Article 49(1) AP I speaks of "violence," a term whose ordinary meaning in the IHL context refers to physical force producing death, injury, damage, or destruction.[83] State practice and expert commentary confirm that the violence threshold excludes non–physical operations: psychological operations and electronic warfare do not qualify as attacks per se because they lack the physical force or destructive consequences that Article 49(1) AP I requires.[84]

This point matters because, in doctrinal terms, "attack" is not a generic label for hostile conduct. It is the jurisdictional trigger for IHL's most demanding precautionary obligations: target verification under Article 57(2)(a)(i) AP I, feasible means selection under Article 57(2)(a)(ii) AP I, and continuing reassessment under Article 57(2)(b) AP I. Cyber operations function through the manipulation of data and logical States, not the direct transfer of kinetic energy. The interpretive question is whether the term "violence" under Article 49(1) AP I must be read mechanistically, requiring physical damage, or functionally, marking the harm threshold that justifies activating the full targeting regime.

The question turns, however, on a prior and unresolved threshold issue of direct doctrinal significance: when does a cyber operation directed

83 AP I, art. 49(1); *see* Int'l Comm. of the Red Cross, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts*, Rep. of the 31st Int'l Conf. of the Red Cross & Red Crescent 37, U.N. Doc. 31IC/11/5.1.2 (2011) [hereinafter ICRC 2011 Challenges Report] ("The term 'acts of violence' denotes physical force."); *see also* Michael N. Schmitt, *'Attack' as a Term of Art in International Law: The Cyber Operations Context*, *in* 4 Proceedings of the International Conference on Cyber Conflict 283, 292 (C. Czosseck et al. eds., 2012) [hereinafter Schmitt, *'Attack' as a Term of Art*].

84 AP I, art. 49(1); 1987 ICRC AP Commentary, *supra* note 18, ¶ 1880, at 603; ICRC 2011 Challenges Report, *supra* note 83, at 37; Tilman Rodenhäuser, *International Humanitarian Law and the Limits of Information or Psychological Operations During Armed Conflicts*, *in* The Oxford Process on International Law Protections in Cyberspace 435, 444–45 (Dapo Akande et al. eds., 2021) [hereinafter Rodenhäuser, *Limits of Info Ops*]; Tallinn Manual 2.0, *supra* note 18, r. 92 cmt. ¶¶ 2–4, 10–13, at 415–19 (defining cyber attack as "a cyber operation, whether offensive or defensive, that is reasonably expected to cause injury or death to persons or damage or destruction to objects"; confirming that "non-violent operations, such as psychological cyber operations and cyber espionage, do not qualify as attacks"; addressing the data deletion question on which the Experts divided, with the majority requiring replacement of physical components and a further minority treating loss of functionality as sufficient irrespective of physical replacement).

against data attain the character of an "attack" within the meaning of Article 49(1) AP I, and does the answer change depending on whether data itself constitutes an "object" under IHL? Where data alteration or deletion produces second-order effects of equivalent gravity to kinetic harm, including the functional disablement of civilian infrastructure or the cascading impairment of systems upon which civilian life materially depends, there is widespread agreement that the functional test for "attack" under Article 49(1) AP I is satisfied, irrespective of the absence of immediate physical damage to tangible property.[85] The antecedent, and as yet unresolved, question whether data itself constitutes an "object" for the purposes of IHL directly conditions the operative scope of two foundational targeting rules: the prohibition on attacking civilian objects under Article 52(1) AP I and the proportionality calculus under Article 51(5)(b) AP I.[86] While this question remains unresolved as a matter of positive international law, with States maintaining that IHL objects must be tangible and Romania asserting the contrary, this analysis proceeds, for the purpose of assessing the structural incompatibility thesis, on the functional approach advanced by the ICRC and a significant body of scholarly opinion, which treats data integral to critical infrastructure functions as within the protective scope of Article 52 AP I; this position is adopted here as the legally more

85 Michael N. Schmitt, *Big Data: International Law Issues During Armed Conflict, in* Big Data and Armed Conflict: Legal Issues Above and Below the Armed Conflict Threshold 149, 163 (Laura A. Dickinson & Edward W. Berg eds., Oxford Univ. Press 2023) [hereinafter Schmitt, Big Data] ("There is widespread agreement that if a cyber operation against data during an armed conflict directly causes the requisite consequences to persons or objects, the operation amounts to an attack. Examples include targeting the data upon which the cyber infrastructure relies in order to function, such as an operation against a reactor's cooling system, causing its meltdown"); Tallinn Manual 2.0, *supra* note 18, r. 92 cmt. ¶ 3; AP I, art. 49(1) ("'Attacks' means acts of violence against the adversary, whether in offence or in defense").

86 Schmitt, Big Data, *supra* note at 163 ("The legal puzzle is how to assess an operation in which the data itself is the object of attack. Alteration, damage, or deletion of data, standing alone, arguably (although not definitively) satisfies the effects requirement for 'attack,' as the data is damaged. This being so, the determinative issue in targeting law is whether data is an 'object' under IHL in the first place"); Tallinn Manual 2.0, *supra* note 18, r. 92 cmt. ¶¶ 10–11 (recording that a majority of the International Group of Experts took the view that "interference with functionality qualifies as damage if restoration of functionality requires replacement of physical components," and that a further sub-group extended this to situations "in which reinstallation of the operating system or of particular data is required in order for the targeted cyber infrastructure to perform the function for which it was designed," such that "if, as a result of a cyber operation deleting or altering data, the infrastructure cannot perform its intended function, the operation in question . . . amounts to an attack"); AP I, arts. 52(1)–(2), 51(5)(b).

defensible interpretation *de lege ferenda* and does not affect the structural incompatibility conclusion, which holds independently of how the data-as-object question is ultimately resolved, because the incompatibility analysis rests on the second-order physical and functional effects of data destruction rather than on data's direct classification as a protected object. If data qualifies as an object, collateral civilian data damage must enter the proportionality equation, and commanders must take precautions in attack in relation to it; if it does not, such harm falls entirely outside the calculus, a result that, as the ICRC has observed, is "difficult to reconcile with the object and purpose of IHL" in an era of pervasive digital dependence.[87] Where, as in autonomous cyber operations of the kind examined in this dissertation, the data targeted is civilian in character, the systemic effects of its destruction may, to the extent the operation satisfies the attack threshold, engage the full targeting regime, including the principles of distinction, proportionality, and precaution under Article 48, 51, and 57 AP I, irrespective of whether any kinetic consequence materializes.[88]

A strictly mechanistic reading offers surface clarity but risks allowing technical form to defeat legal function: foreseeable civilian suffering becomes doctrinally peripheral merely because it is mediated by code rather than by explosive force. A functional reading, by contrast, treats

87 Schmitt, Big Data, *supra* note 85, at 164 ("In contrast, if data is not an object as a matter of IHL, conducting cyber operations intended to alter or delete civilian data is lawful so long as there are no damaging second-order effects to civilian cyberinfrastructure or other systems, or harm to civilians"; "The matter is unsettled and contentious"); *id.* at 164–65 (documenting State divergence on whether data constitutes a protected object under the law of armed conflict, with one State maintaining that "objects for the purposes of LOAC have always been understood to be tangible things and this understanding is not domain-specific" and that "only tangible things can constitute objects," while another asserts that "cyber operations against data do trigger the application of IHL" and "cannot be directed against those data that represent a civilian object which must be protected under the principle of distinction"); Int'l Comm. of the Red Cross, *International Humanitarian Law and Cyber Operations During Armed Conflicts*, ICRC Position Paper, at 8 (Nov. 2019) ("the assertion that deleting or tampering with such essential civilian data would not be prohibited by IHL in today's data-reliant world seems difficult to reconcile with the object and purpose of IHL"); AP I, arts. 52(1)–(2).

88 AP I, arts. 48, 51(4)(b)–(c), 57(2)(a)(ii)–(iii); Tallinn Manual 2.0, *supra* note 18, r. 92 cmt. ¶¶ 13–14 (recording that a majority of the International Group of Experts took the position that "although there might be logic in characterising" large-scale disruptive data operations "as an attack, the law of armed conflict does not presently extend this far," and that "not all cyber operations qualify as attacks," confirming the conditionality of the full targeting regime on satisfaction of the harm threshold).

"violence" as a limiting concept that preserves the gatekeeping function of "attack" while remaining responsive to the humanitarian object and purpose of the targeting rules, an interpretive approach required by Article 31(1) VCLT and embedded in AP I's protective architecture.[89] On this view, "violence" marks the point at which the expected consequences engage the protected interests of persons and objects in a manner sufficient in gravity to justify the activation of IHL's strictest precautionary obligations, regardless of the physical mechanism by which those consequences are produced, thereby justifying activation of the strict precautionary regime. This reading does not erase the boundary between attacks and non–attacks; it insists that the boundary be drawn with reference to expected humanitarian risk rather than to the physicality of the immediate mechanism.

For autonomous cyber weapons, the interpretive stakes increase. Autonomy reduces the temporal and epistemic space in which a human decision–maker can evaluate consequences, thereby increasing the importance of ex ante legal characterization and design–stage constraint. If a capability is designed to select targets, propagate, and generate effects without real–time human authorization, then whether it performs "attacks" is not an after–the–fact label; it is a legality condition that must be resolved upstream in the architecture of authorization and Article 36 AP I review. State practice demonstrates that this functional reading can be operationalized through bounded threshold criteria that preserve essential peacetime/wartime distinctions while extending IHL protections to effects–based harm.

89 Vienna Convention on the Law of Treaties art. 31(1), *opened for signature* May 23, 1969, 1155 U.N.T.S. 331; AP I, arts. 49(1), 57(2)(a)–(b); Int'l Law Comm'n, *Fragmentation of International Law: Difficulties Arising from the Diversification and Expansion of International Law*, U.N. Doc. A/CN.4/L.682, ¶¶ 37, 413 (Apr. 13, 2006) (finalized by Martti Koskenniemi) [hereinafter ILC Fragmentation Report], https://legal.un.org/ilc/documentation/english/a_cn4_l682.pdf (last visited Apr. 16, 2026); Campbell McLachlan, *The Principle of Systemic Integration and Article 31(3)(c) of the Vienna Convention*, 54 Int'l & Comp. L.Q. 279, 279–80 (2005) [hereinafter McLachlan, *Systemic Integration*].

2. State Practice Convergence on a Bounded Functional Harm Test

The Tallinn Manual 2.0 experts' concern that a purely functional approach would collapse essential distinctions between lawful cyber espionage and unlawful armed attack is doctrinally sound and requires a direct response before the bounded functional harm test advanced in this section can be accepted. Specifically, if functional degradation of military networks suffices to constitute an "attack" under Article 49 AP I, States conducting cyber intelligence operations against adversary military infrastructure would face armed attack allegations under the law of armed conflict, a consequence that would destabilize the peacetime/wartime distinction upon which the entire *jus in bello* framework rests.[90]

The State practice documented above, particularly France's functional harm threshold under Article 49 AP I,[91] the Netherlands' effects–based *jus contra bellum* assessment,[92] and the U.S. contextual approach emphasizing consequences functionally equivalent to those produced by kinetic force,[93] demonstrate that effects–based approaches need not collapse peacetime/wartime distinctions when bounded by three cumulative thresholds derived from IHL's core principles.

The United States approach applies this proportionality standard to cyber operations through effects–based equivalence, holding that cyber operations producing effects "that, if caused by traditional physical means, would be regarded as a use of force" are subject to the same law of war requirements, including the obligation to refrain from attacks

90 Tallinn Manual 2.0, *supra* note 18, r. 92 cmt. ¶¶ 12–14.

91 Ministère des Armées (Fr.), *Droit international appliqué aux opérations dans le cyberespace* [International Law Applied to Operations in Cyberspace] 13–14 (Sept. 9, 2019), https://www.defense.gouv.fr/sites/default/files/ministere-armees/Droit%20international%20appliqu%C3%A9%20aux%20op%C3%A9rations%20dans%20le%20cyberespace.pdf (last visited Apr. 16, 2026) [hereinafter French Ministry Cyber Doctrine].

92 Letter from the Minister of Foreign Affairs and the Minister of Defense (Neth.) to the President of the House of Representatives on the International Legal Order in Cyberspace, Appendix to Parliamentary Papers II 2019/20, 26 643, No. 570, at 3 (July 5, 2019) [hereinafter Netherlands Cyberspace Letter 2019].

93 AP I, art. 51(5)(b); DoD Law of War Manual, *supra* note 67, § 2.4.1.2, at 61.

where expected civilian harm would be excessive in relation to anticipated military advantage.[94]

The United Kingdom's position reinforces this functional approach. In its 2021 Statement to the UN Group of Governmental Experts, the UK Government confirmed that "a cyber operation is capable of being an 'attack' under IHL where it has the same or similar effects to kinetic action that would constitute an attack," emphasizing that IHL principles of "distinction, proportionality, humanity and military necessity apply in the same way as they do to an attack by any other means."[95] This effects–based test supports the application of IHL protections to cyber operations that produce effects equivalent to "acts of violence" under the definition of attacks under Article 49(1) AP I, thereby implicating the protection of civilian objects under Article 52 AP I and the precautionary obligations under Article 57 AP I when cyber operations target infrastructure with potential civilian impact.[96]

Germany's position emphasizes the necessity of meaningful human control over autonomous weapons systems, particularly in cyber operations where effects may cascade unpredictably. In its 2021 joint submission with France to the CCW Group of Governmental Experts (CCW/GGE.1/2021/WP.5), Germany advocated a "clear commitment by High Contracting Parties not to develop, produce, acquire, deploy or use fully autonomous lethal weapons systems operating completely outside human chain of command and control."[97] Applied to autonomous cyber weapons, this position establishes that systems precluding commanders from exercising "judgement with regard to

94 AP I, art. 51(5)(b); DoD Law of War Manual, *supra* note 67, § 2.4.1.2, at 61.

95 United Kingdom, *Application of International Law to States' Conduct in Cyberspace*, Statement to the U.N. GGE ¶ 24 (2021), https://assets.publishing.service.gov.uk/media/60b775388fa8f54899011dec/application-of-international-law-to-states-conduct-in-cyberspace-uk-statement.pdf (last visited Apr. 16, 2026) [hereinafter UK GGE Statement 2021].

96 AP I, arts. 49(1), 51(2), 52, 57(2)(a); *see also* ICRC Customary IHL Study, *supra* note 18, rr. 7, 14, 15.

97France & Germany, *Outline for a Normative and Operational Framework on Emerging Technologies in the Area of LAWS* § 2(b), U.N. Doc. CCW/GGE.1/2021/WP.5 (Sept. 27, 2021), https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Seventh_Group_of_Governmental_Experts(2021)/CCW-GGE.1-2021-WP.5_English.pdf (last visited Apr. 16, 2026) [hereinafter France & Germany, *LAWS Framework*].

compliance with IHL in the framework and context of an attack" constitute structural IHL violations under Article 57's precautionary requirements. The operational framework further requires that "humans must have sufficient assurance that weapons systems, once activated, act in a foreseeable manner" and that "commanders and operators must have a sufficient understanding of the weapons systems' way of operating, which enables the commanders and operators to predict (prospective focus) and explain (retrospective) the behavior of the weapons systems."[98] Germany's formulation thus supplies a State-practice anchor for the prospective-and-retrospective explicability standard: a weapon whose targeting outputs a commander cannot predict prospectively or explain retrospectively fails the Article 57 feasibility threshold as a matter of existing treaty obligation, not merely as a proposed regulatory development.

The following section demonstrates, at the level of each individual targeting obligation, precisely how and why those duties cannot be satisfied.

3. The Stuxnet Precedent: Precision Design and Indiscriminate Propagation

Physical destruction of objects, whether accomplished through kinetic or cyber means, constitutes an attack subject to targeting restrictions. The widely cited Stuxnet operation against Iranian nuclear enrichment centrifuges provides the paradigmatic illustration: between late 2009 and early 2010, Iran decommissioned and replaced on the order of 1,000 IR–1 centrifuges at Natanz, which technical analysis by the Institute for Science and International Security reasonably attributes to Stuxnet's

98 *Id.* § 2(c) (sixth bullet); *id.* § 3(b); AP I, arts. 36, 57.

destructive effects, thereby satisfying the violence threshold under Article 49(1) AP I.[99]

The deeper doctrinal value of Stuxnet for this chapter, however, lies not in the baseline proposition that cyber means can produce attacks, but in the relationship between design intent, propagation behavior, and the feasibility of IHL precautions. Stuxnet was engineered for intended precision: its activation logic was keyed to a highly specific Siemens programmable logic controller configuration of the type used at facilities like Natanz, suggesting that its designers attempted to satisfy Article 57(2)(a)(i)'s verification obligation by limiting effects to confirmed military objectives. This pattern-discriminating activation logic, paired with indiscriminate propagation mechanisms, illustrates what might be termed the *cyber precision paradox*: technical constraints designed to satisfy IHL precautions at the design phase are defeated by propagation behavior at the deployment phase, rendering Article 57's verification obligations structurally infeasible for autonomous cyber weapons.[100] The ICRC has confirmed, as recently as April 2025, that AI-enabled autonomous cyber capabilities could increase the risks of indiscriminate attacks, incidental civilian harm, including damage to critical civilian infrastructure, and uncontrolled escalation of conflict in complex and interconnected digital environments.[101] States, at the 34th International Conference of the Red Cross and Red Crescent, expressly recognized that the use of artificial intelligence and other emerging

99 AP I, art. 49(1); David Albright & Paul Brannan, *Did Stuxnet Take Out 1,000 Centrifuges at the Natanz Enrichment Plant?*, Inst. for Sci. & Int'l Sec. 1 (Dec. 22, 2010), https://isis-online.org/isis-reports/did-stuxnet-take-out-1000-centrifuges-at-the-natanz-enrichment-plant/ (last visited Apr. 16, 2026) (reporting that "Iran decommissioned and replaced about 1,000 IR-1 centrifuges" at Natanz in late 2009–early 2010 and concluding that Stuxnet infection "could have resulted" in this breakage) [hereinafter Albright & Brannan, *Stuxnet Centrifuges*]; Tallinn Manual 2.0, *supra* note 18, r. 92 & cmt. ¶¶ 3–4.

100 AP I, art. 57(2)(a)(i); Falliere, Murchu & Chien, *supra* note 77, at 2, 39, 42 (S7-315-2 PLCs; 807–1210 Hz); *id.* at 5 (≈100,000 hosts); *id.* at 7 ("collateral damage"); Albright & Brannan, Stuxnet Centrifuges, *supra* note 99, at 1.

101 Int'l Comm. of the Red Cross, *Submission to the United Nations Secretary-General on Artificial Intelligence in the Military Domain*, ODA/2025-00029/AIMD, at 6 (Apr. 11, 2025), https://www.icrc.org/sites/default/files/2025-04/ICRC_Report_Submission_to_UNSG_on_AI_in_military_domain.pdf [hereinafter ICRC, AI Military Domain Submission (2025)]; *see also* Kubo Mačák, *Artificial Intelligence in Armed Conflict: CyCon 2025 Series, Introduction*, Article of War (Lieber Inst. W. Point, Aug. 18, 2025), https://lieber.westpoint.edu/artificial-intelligence-armed-conflict-cycon-2025-series-introduction [hereinafter Mačák, *CyCon Introduction*] (summarizing the same ICRC findings).

technologies in malicious ICT activities may further increase their scale and speed, as well as the harm they may cause.[102]

This paradox maps directly onto Article 57's structure. IHL does not regulate only the terminal effect; it regulates the decision to employ a means of warfare whose foreseeable operational behavior creates civilian risk. Article 57(2)(b) requires commanders to "cancel or suspend an attack if it becomes apparent that the objective is not a military one" or that harm would be excessive.[103] Where a cyber capability lacks effective recall, termination, or suspension mechanisms once deployed, this continuing obligation is placed under structural strain.[104]

Military operational planning timelines illuminate the distinction between verification difficulty and *verification impossibility*. Traditional kinetic targeting cycles employ the F2T2EA methodology, Find, Fix, Track, Target, Engage, Assess, which allocates hours to days for intelligence collection, target development, and legal review before strike authorization; deliberate targeting typically operates on planning horizons measured in hours to days, with planning cycles for deliberate targets that may extend to 72–96 hours under U.S. doctrine to permit

---

102 34th International Conference of the Red Cross and Red Crescent, Resolution 2: *Protecting Civilians and Other Protected Persons and Objects Against the Potential Human Cost of ICT Activities During Armed Conflict*, Doc. 34IC/24/R2, preambular ¶ 11 (Oct. 2024), https://rcrcconference.org/app/uploads/2024/10/34IC_R2-ICT-EN.pdf [hereinafter 34th ICRC Conference, Res. 2]; Mačák, CyCon Introduction, *supra* note 101 (citing *id.*).

103 AP I, art. 57(2)(b) ("An attack shall be cancelled or suspended if it becomes apparent that the objective is not a military one or is subject to special protection or that the attack may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated."); 1987 ICRC AP Commentary, *supra* note 18, ¶ 2220 (confirming that the cancellation and suspension obligation "applies not only to those planning or deciding upon attacks, but also and primarily, to those executing them"); *id.* ¶ 2221 (interpreting "if it becomes apparent" as requiring the attacker to abstain from attack upon receipt of any new information indicating that the intended objective is not a military one, irrespective of prior orders or planning-phase assessments).

104 AP I, art. 57(2)(b) ("an attack shall be cancelled or suspended if it becomes apparent that the objective is not a military one . . . or that the attack may be expected to cause incidental loss of civilian life . . . which would be excessive in relation to the concrete and direct military advantage anticipated"); 1987 ICRC AP Commentary, *supra* note 18, ¶ 2220 (confirming that the cancellation and suspension obligation "applies not only to those planning or deciding upon attacks, but also and primarily, to those executing them" and that the duty is a continuing one operative throughout execution); *id.* ¶ 2221 (requiring abstention from attack upon receipt of any new information indicating the intended objective is not a military one; "even greater caution is required" where the attacker lacks direct observation of the objective); Falliere, Murchu & Chien, *supra* note 77, at 2–3, 21–23 (documenting that Stuxnet's peer-to-peer propagation architecture and air-gapped deployment environment structurally prevented command-and-control termination once the payload was activated, rendering any post-deployment suspension mechanism operationally unavailable).

collateral damage estimation, proportionality assessment, and integration of component requirements mandated by doctrine and policy."[105] The proportionality test, whether expected incidental civilian harm is excessive in relation to the concrete and direct military advantage anticipated, must be applied on the basis of information reasonably available at the time of the attack, requiring commanders to make good-faith assessments in circumstances where the answers to specific questions depend on the circumstances appraised in the light of the humanitarian ends that justify the restraints, and where determining the proper relation between means and ends in situations of great complexity and uncertainty is never easy."[106] The constraint is not merely procedural: the law of armed conflict recognizes that proportionality determinations involve highly open-ended inquiries whose answers may remain subjective and imprecise, and that the "reasonable military commander" standard cannot be reduced to a uniform calculus precisely because commanders with different doctrinal backgrounds, degrees of combat experience, and national military histories would not always agree in close cases, though they would agree where injury to noncombatants was clearly disproportionate to the

105 Joint Chiefs of Staff, *Joint Publication 3-60, Joint Targeting* (Jan. 31, 2013), https://jfsc.ndu.edu/Portals/72/Documents/JC2IOS/Additional_Reading/1F4_jp3-60.pdf (last visited Apr. 16, 2026), at II-23 fig. II-10, II-28–II-30 figs. II-15 to II-16, III-2 to III-3 (detailing the six steps of F2T2EA dynamic targeting, including the Target and Engage steps at figures II-15 and II-16; mandating that the Staff Judge Advocate review target selection and force assignment for legal compliance prior to the release of tasking orders in Phase 4, before the Engage step in Phase 5).

106 AP I, art. 57(2)(a)(i) (imposing on those who plan or decide upon an attack the obligation to "do everything feasible to verify" that the objectives to be attacked are military objectives, with feasibility assessed in light of the "circumstances ruling at the time"); U.S. Dep't of Defense, *Law of War Manual* § 5.12.1, at 269–70 (rev. ed. July 31, 2023) (confirming that proportionality requires assessment of the totality of expected incidental civilian harm against the concrete and direct military advantage anticipated in the circumstances ruling at the time); Ian Henderson & Kate Reece, *Proportionality Under International Humanitarian Law: The "Reasonable Military Commander" Standard and Reverberating Effects*, 51 Vand. J. Transnat'l L. 835, 840 (2018) (quoting Oscar Schachter, Remarks, 86 Am. Soc'y Int'l L. Proc. 39, 39 (1992) ("the answers to specific questions depend on the circumstances, appraised in the light of the humanitarian ends that justify the restraints"; "determining the proper relation between means and ends in situations of great complexity and uncertainty is never easy")).

military advantage gained."[107] Even within dynamic targeting prosecuted against fleeting targets, the Target step mandates coordination and legal assessment before the Engage step authorizes weapon release, preserving the commander's irreducible duty to conduct situationally contingent ethical deliberation under compressed timelines.[108] These structural features of targeting doctrine, temporal discipline, non-quantifiable balancing, and mandatory human judgment establish that verification of compliance with distinction and proportionality principles during autonomous weapon system operation is not merely difficult but conceptually impossible absent continuous human evaluative control at decision-critical nodes.

Autonomous cyber weapons compromise the precautionary requirements through temporal compression at three stages. First, pre-positioned malware autonomously activated upon detecting target configurations during armed conflict may propagate through network infrastructure faster than human operators can execute the verification cycle required for Article 57(2)(a)(i) compliance. Second, machine-speed autonomous propagation may generate cascade effects that outpace the timeframe within which commanders can assess proportionality under Article 51(5)(b), rendering feasible precautions structurally impossible rather than merely difficult.

---

107 Office of the Prosecutor, Int'l Criminal Tribunal for the Former Yugoslavia, *Final Report to the Prosecutor by the Committee Established to Review the NATO Bombing Campaign Against the Federal Republic of Yugoslavia* ¶¶ 49–50, 39 I.L.M. 1257, 1280–81 (2000) ("The answers to these questions are not simple. It may be necessary to resolve them on a case-by-case basis, and the answers may differ depending on the background and values of the decision maker. It is unlikely that a human rights lawyer and an experienced combat commander would assign the same relative values to military advantage and to injury to noncombatants. . . . There will be many cases where reasonable military commanders will agree that the injury to noncombatants or the damage to civilian objects was clearly disproportionate to the military advantage gained."); Henderson & Reece, *supra* note 106, at 840–42.

108 AP I, arts. 51(5)(b), 57(2)(a)(iii), 57(2)(b); Joint Pub. 3-60, *supra* note 105, at II-23 fig. II-10, II-28–II-30 figs. II-15 to II-16 (setting out the six steps of dynamic targeting—Find, Fix, Track, Target, Engage, Assess—and locating the commander's decision and Engage steps within Phase 5 of the joint targeting cycle); U.S. Dep't of Defense, DoD Instruction 3000.17: *Civilian Harm Mitigation and Response* § 1.2(c), at 5 (Dec. 21, 2023), https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/300017p.pdf ("Commanders at all levels have a great responsibility to exercise the leadership necessary to reduce the risk of harm to civilians and civilian objects during military operations involving the use of force."); U.S. Dep't of Defense, *Law of War Manual* §§ 5.12.2–5.12.3, at 272–74 (rev. ed. July 31, 2023) (clarifying that proportionality encompasses all reasonably expected incidental civilian harm, including foreseeable reverberating effects, and must be assessed ex ante in light of the circumstances ruling at the time); Henderson & Reece, *supra* note 106, at 840–42.

Third, where network conditions change between implant deployment as a peacetime intelligence operation and autonomous activation as a wartime attack, commanders lack contemporaneous access to "circumstances ruling at the time" necessary for Article 57(2)(a) assessment, rendering verification temporally impossible even where technically feasible.[109]

Stuxnet's architecture lacked dynamic termination mechanisms, rendering mid-campaign suspension technically impossible once propagation exceeded the originally seeded Iranian organizational domains, a design failure that directly instantiated the third temporal compression problem identified above. Autonomous weapon systems incorporating machine learning pose an acute challenge to the prohibition on weapons that are by nature indiscriminate under Customary IHL Rule 71. Where an AWS functions through opaque algorithmic processes that preclude the human operator from understanding, predicting, or explaining its targeting outputs, the weapon becomes tantamount to blind firing – a violation of the requirement to direct attacks at specific military objectives. Machine learning–based AWS exemplifies this risk: their functioning can change after activation such that force is applied in circumstances unforeseen by the human user, effectively severing the causal link between the commander's legal assessment at activation and the system's eventual

109 AP I, arts. 51(5)(b), 57(2)(a)(i); U.S. Dep't of Defense, *Law of War Manual* § 5.2.3.2, at 194 (rev. ed. July 31, 2023) ("'Feasible precautions' are those that are practicable or practically possible, taking into account all circumstances ruling at the time, including humanitarian and military considerations."); 1987 ICRC AP Commentary, *supra* note 18, ¶ 2195 (explaining that the "everything feasible" standard in article 57 codifies and integrates existing precautionary obligations in attack); Nat'l Cybersecurity & Commc'ns Integration Ctr., U.S. Dep't of Homeland Sec., *Technical Alert TA17-181A: Petya Ransomware* (July 1, 2017, last revised Feb. 15, 2018), https://www.cisa.gov/news-events/alerts/2017/07/01/petya-ransomware (documenting fully automated EternalBlue-based lateral propagation scanning network subnets without human intervention at each exploitation step); Do Duc Anh Nguyen et al., *How Fast Does Malware Leveraging EternalBlue Propagate? The Case of WannaCry and NotPetya*, *in* Proceedings of the IEEE 10th International Conference on Network Softwarization (NetSoft 2024) 1, 4, https://hal.science/hal-04645862v1 (finding in controlled experiments that approximately 20% of a 50-host network could be compromised in under 50 seconds, with up to 16 hosts infected per 100-second interval); Int'l Comm. of the Red Cross, *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons* 9–10 (2016), https://www.icrc.org/en/doc/assets/files/2016/icrc-report-autonomous-weapons.pdf ("Not knowing when, how, and where a machine would carry out an attack would prevent the user, or commander, from being able to implement his/her legal obligations with respect to the conduct of hostilities.").

target selection. The ICRC concludes that such systems "would likely be indiscriminate by nature" because commanders cannot reasonably anticipate their effects, thereby failing to direct or limit the weapon's effects as AP I, Article 51(4) requires. This doctrinal position, grounded in the GGE's 2019 recognition that machine learning introduces "a risk of unpredictability," supports a categorical prohibition on deploying AWS, whose opacity renders pre–strike IHL assessments impossible.[110]

4. Functional Harm Threshold: State Practice Survey (2019–2025)

The July 2025 UN Open–Ended Working Group on ICTs (2021–2025), open to all 193 UN Member States, reaffirmed the "cumulative and evolving framework for responsible State behavior in the use of ICTs" and confirmed that "international law, in particular the Charter of the United Nations, is applicable and essential to maintaining peace, security and stability in the ICT environment" and that "existing obligations under international law are applicable to States' ICT–related activity."[111] Within that normative environment, the functional harm threshold

articulated here is deliberately modest: not an unqualified redefinition of attack, but a constrained functional harm test calibrated to preserve the armed–conflict/peacetime boundary while preventing IHL evasion through technical modality.

The bounded threshold contains three cumulative elements: (i) loss of civilian capability at a scale implicating protected persons or objects;

110 ICRC, *AWS Selected Issues*, *supra* note 15, at 7; Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, *Report of the 2021 Session*, ¶ 25(c), U.N. Doc. CCW/GGE.1/2021/3 (Feb. 22, 2022) [hereinafter CCW GGE 2021 Session Report], https://undocs.org/CCW/GGE.1/2021/3 (last visited Apr. 16, 2026) (recommending prohibition where incidental civilian harm "cannot be reasonably foreseen or are not fully understood by a human operator"); Jean-Marie Henckaerts & Louise Doswald-Beck, Customary International Humanitarian Law r. 71 (Cambridge Univ. Press 2005); AP I, art. 51(4)(b).

111 U.N. Open-Ended Working Group on Security of and in the Use of Information and Communications Technologies 2021–2025, *Final Report*, annex, ¶¶ 5, 41(a), U.N. Doc. A/80/257 (July 11, 2025), https://undocs.org/A/80/257 (last visited Apr. 16, 2026) [hereinafter OEWG Final Report 2025].

(ii) effects of sufficient humanitarian gravity — measured by severity, duration, and scale of civilian suffering — to engage the protective purposes of the targeting regime under Articles 48–57 AP I, irrespective of the physical mechanism by which those effects are produced; and (iii) absence of clear military advantage justification proportionate to the civilian disruption.

NotPetya, analysed in full in Section II.3 below, provides the paradigmatic illustration of all three cumulative elements. The Tallinn Manual experts' concern that a purely functional approach would collapse essential peacetime/wartime distinctions is addressed directly by the bounded character of this threshold: the three cumulative elements exclude routine network disruptions, cyber espionage, and transient service interruptions, while capturing operations whose humanitarian effects are functionally equivalent to kinetic attack.

France's functional formulation "service no longer provided; intervention required" identifies a threshold that excludes routine network frictions and transient slowdowns while capturing sustained loss of operational function.[112] The United Kingdom confirmed in its 2021 statement to the UN Group of Governmental Experts that a cyber operation constitutes an "attack" under IHL where it has "the same or similar effects to kinetic action that would constitute an attack," and that the principles of distinction, proportionality, humanity, and military necessity apply to such operations in the same way they apply to kinetic attacks — a position that treats effects-based equivalence, not physical mechanism, as the operative legal criterion. The U.S. contextual effects analysis similarly resists categorical inflation while allowing functional equivalence where consequences mirror those of kinetic attacks in humanitarian terms. Rather than adopting bright–line rules predicated solely on loss of functionality or service disruption, the U.S. Department

112 French Ministry Cyber Doctrine, *supra* note 91, at 14.

of Defense Law of War Manual employs a flexible, effects–based inquiry that examines whether the humanitarian impact of a cyber operation — measured by its severity, scale, duration, and consequences for the civilian population — rises to the gravity that the law of armed conflict's attack rules were designed to regulate. This approach distinguishes between cyber operations that "destroy enemy computer systems" (constituting "attacks" subject to IHL's full panoply of protections) and those that merely "deface a government webpage," cause "a minor, brief disruption of internet services," or "briefly disrupt, disable, or interfere with communications and disseminate propaganda" (which are "not attacks" and thus theoretically "may be directed at civilians or civilian objects"). Crucially, however, even sub–attack cyber operations "should not be conducted in a way that unnecessarily causes inconvenience to civilians or neutral persons," reflecting an implicit precautionary obligation grounded in the principles of humanity and military necessity.[113] The ICRC's loss–of–functionality approach anchors the attack definition in harm to persons or objects (including disabling cyber infrastructure), thereby preserving the humanitarian purpose of the targeting rules while ensuring that operations rendering civilian networks dysfunctional receive IHL's most detailed protections.[114]

For autonomous cyber weapons, the functional harm threshold has direct operational implications. The more a capability is designed to generate widespread functional loss across interconnected infrastructure, whether through propagation logic, autonomous target selection mechanisms, or foreseeable reverberating effects, the more

113 DoD Law of War Manual, *supra* note 67, § 16.5.1, at 1034; *id.* § 16.5.2, at 1035.

114 ICRC Cyber Operations Report, *supra* note 23, at 7–8.

likely it is to meet the Article 49(1) threshold.[115] The significance of this conclusion for Tier 3 autonomous cyber weapons is compounded by temporal disaggregation: where deployment authorization precedes activation by months or years, the effects that trigger the attack threshold materialize in circumstances that no human decision-maker contemporaneously assessed against the humanitarian criteria Article 49(1) presupposes. This temporal dimension is not incidental to the threshold analysis — it is the mechanism by which Tier 3 architecture structurally defeats the compliance preconditions that the threshold exists to protect. Conversely, the more uncertain the civilian impact and the more opaque the propagation pathways, the more urgently IHL requires that any residual uncertainty be allocated away from civilians. Where propagation uncertainty creates doubt about which objects will be affected, whether military networks or civilian infrastructure, the *in dubio pro civili* principle — the obligation to resolve doubt in favor of civilian protection — under Article 52(3) AP I requires presuming that uncertain targets are civilian objects protected from attack.[116]

Autonomous target–selection systems fundamentally complicate the Article 57 verification obligation because they operationalize distinction through probabilistic confidence thresholds rather than categorical legal determinations that an object satisfies the cumulative test under Article 52(2) AP I for military–objective status. This characterization requires direct engagement with the limiting interpretation advanced by Schmitt and Boothby, who argue that Article 52(3)'s presumption of civilian status applies only to objects "normally dedicated to civilian purposes" — that is, objects with a documented civilian-use history — and that its

115 AP I, art. 49(1); ICRC Cyber Operations Report, *supra* note 23, at 7–8; Kubo Mačák & Ewan Lawson, *Avoiding Civilian Harm During Military Cyber Operations: Six Key Takeaways*, ICRC Humanitarian L. & Pol'y Blog (June 15, 2021), https://blogs.icrc.org/law-and-policy/2021/06/15/avoiding-civilian-harm-military-cyber-operations/ (last visited Apr. 16, 2026) [hereinafter ICRC, *Avoiding Civilian Harm*]; *see also* Rain Liivoja & Ann Väljataga eds., Autonomous Cyber Capabilities Under International Law 45–46 (NATO CCDCOE 2021) [hereinafter Liivoja & Väljataga, *Autonomous Cyber Capabilities*].

116 AP I, art. 52(3) ("In case of doubt whether an object which is normally dedicated to civilian purposes . . . is being used to make an effective contribution to military action, it shall be presumed not to be so used."); ICRC Cyber Operations Report, *supra* note 23, at 7–8; Liivoja & Väljataga, Autonomous Cyber Capabilities, supra note 259, at 45–46.

scope is accordingly limited in dual-use and military-infrastructure contexts where no such history is established. That interpretation finds textual support in the phrase "normally dedicated to civilian purposes." Three responses sustain the broader reading advanced here. First, the ICRC Commentary to AP I, Article 52, ¶ 2024 (1987, updated 2016) does not condition the presumption's application on proof of a documented civilian-use record; it extends the presumption to any object in respect of which doubt as to military use exists at the time of attack. Second, the protective purpose of Article 52(3), construed under Article 31(1) VCLT, requires extending the presumption precisely to cases where the attacking party's own targeting system cannot resolve the ambiguity — algorithmic uncertainty is the paradigm case of the doubt the drafters intended to resolve in favor of civilian protection. Third, Schmitt's limiting interpretation would systematically exclude network infrastructure, the primary operational environment of autonomous cyber weapons, from the protection under Article 52(3) AP I, since most network nodes lack a "normally dedicated to civilian purposes" documentation record, a result that is inconsistent with IHL's object and purpose in an age of pervasive digital infrastructure. Accordingly, where an autonomous weapon's probabilistic classification generates residual uncertainty about an object's military status, thereby triggering the mandatory civilian presumption, that object "shall be presumed not to be so used" unless "it is certain" that it serves military functions. The ICRC Commentary confirms that "in case of doubt, the safety of the civilian population, which is the aim of the Protocol, must be taken into consideration," and that commanders "ordering or executing the attack must have sufficient information available" to satisfy the requirement that destruction offer "a definite military advantage in the circumstances ruling at the time." Absent human verification capable of dispelling algorithmic uncertainty through evaluation of "all circumstances ruling at the time, including

humanitarian and military considerations," authorizing the strike would violate the precautionary duty to "do everything feasible to verify that the objectives to be attacked are neither civilians nor civilian objects... but are military objectives."[117] The bounded functional harm threshold thereby generates a cascading legal consequence: once an autonomous cyber operation crosses the Article 49(1) attack threshold, every subsequent obligation in Articles 48 through 58 of AP I engages — and each of those obligations, as Section II demonstrates, presupposes precisely the human decision-maker that Tier 3 architecture has eliminated.

Where verification is structurally infeasible, the characterization of a cyber capability as an attack-generating weapon is itself a preventive governance trigger: Article 36 AP I review must determine whether its employment can ever comply with Articles 48, 51, 52, and 57, or whether its design precludes lawful use.[118]

The qualification of autonomous cyber weapon outputs as "attacks" under Article 49(1) — established in the preceding section through convergent State practice and the ICRC's functional harm threshold — has a determinate doctrinal consequence: it engages the full protective architecture of AP I Articles 48 through 58. What remains to be demonstrated is that this engagement is not merely formal. The rules engaged are structurally incompatible with the weapons they now govern because those rules presuppose epistemic and operational conditions that autonomous cyber weapons architecturally eliminate. It is that demonstration — carried across distinction, proportionality, and precaution — to which Section II now turns.

117 AP I, arts. 52(2)–(3), 57(2)(a)(i), 57(2)(a)(iii); 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 2022, 2024, 2034.

118 *Id.* arts. 36, 48, 51, 52, 57.

II. Distinction, Proportionality, and Precaution: Structural Incompatibility with Core IHL Targeting Obligations

This Section makes one argument: the three core rules of IHL targeting — distinction, proportionality, and precaution — presuppose a human being present at the moment of attack who can apply them. Tier 3 autonomous cyber weapons eliminate that human. The rules therefore cannot operate. What follows proves that claim, rule by rule.

The structural incompatibility demonstrated in this Section is confined to systems meeting the Chapter 1 definition of Tier 3 autonomous cyber weapons: systems whose targeting outputs are generated by opaque machine-learning inference architectures, specifically deep neural networks or reinforcement learning models operating in adversarial environments, under conditions of temporal disaggregation between access authorization and effects generation, and without architecture-level human interrupt capability at the moment of target selection. Systems retaining a human-in-the-loop at the target-selection node, including remotely piloted systems and human-supervised semi-autonomous systems, fall outside this dissertation. The incompatibility thesis is categorical with respect to Tier 3 systems and conditional with respect to Tier 2 systems, depending on the degree of retained human oversight. This scope limitation is not a concession: it is the doctrinal precision that distinguishes a structural prohibition argument from a categorical ban proposal, and it is what renders the incompatibility thesis consistent with the bounded *de lege lata* claim mapped in the Chapter's opening normative-tier statement.

The three core targeting obligations of international humanitarian law — distinction, proportionality, and precaution — do not operate as independent rules susceptible to isolated compliance assessments; they constitute an integrated normative architecture in which each obligation presupposes the epistemic and operational conditions that the others

require. Distinction provides the threshold legal judgment that an object qualifies as a military objective under Article 52(2); proportionality calibrates the permissible incidental harm against the concrete military advantage that objective's neutralization offers under Article 51(5)(b); and precaution operationalizes both through the affirmative verification and cancellation duties of Article 57 — duties that must be discharged at the operative moment of attack. This tripartite architecture rests on a shared foundational presupposition: that a human decision–maker, possessing adequate contemporaneous intelligence, stands at the moment of attack with the cognitive capacity and technical ability to integrate legal judgment across all three obligations simultaneously. Autonomous cyber weapons defeat that presupposition not at the margins of one rule but structurally and cumulatively — distinction is defeated because machine learning classifiers cannot, as this analysis demonstrates, perform the semantic functional judgment that Article 52(2) AP I requires; proportionality is structurally unavailable because network topology and cascade propagation dynamics render the expected civilian harm of a given operation objectively unquantifiable under any planning methodology available at the moment of authorization; and precaution is architecturally foreclosed because temporal disaggregation and algorithmic opacity eliminate both the verification capacity and the cancellation mechanism that Article 57 presupposes — conclusions grounded, as the sections below demonstrate, in the treaty text's own epistemic and temporal presuppositions, not in a policy preference for restrictive regulation.

1. Distinction: The Classification Incapacity of Algorithmic Targeting

The principle of distinction, codified in Article 48 of AP I, requires parties to distinguish between civilians and combatants and between civilian objects and military objectives and to direct operations only against military objectives.[119] The ICJ characterized distinction as one of the "cardinal principles" of IHL, constituting "intransgressible principles of international customary law."[120] The ICRC Customary IHL Study confirms that distinction binds all parties in both international and non–international armed conflicts as Rule 1 of customary law.[121]

The customary international humanitarian law status of the distinction principle extends beyond general obligations to specific operational requirements. Rule 1 establishes that "the parties to the conflict must at all times distinguish between civilians and combatants," while Rule 7 requires that "the parties to the conflict must at all times distinguish between civilian objects and military objectives."[122] These rules apply in both international and non–international armed conflicts, extending the distinction obligation to cyber operations conducted by non–State armed groups and States not party to AP I.[123] Rule 11 of the ICRC Customary IHL Study prohibits indiscriminate attacks in both international and non–international armed conflicts; Rule 12 defines them to include, inter alia, attacks that "employ a method or means of combat the effects of which cannot be limited as required by

119 AP I, art. 48.

120 Nuclear Weapons Advisory Opinion, *supra* note 39, ¶¶ 78–79 ("The cardinal principles contained in the texts constituting the fabric of humanitarian law are the following" and "they constitute intransgressible principles of international customary law.").

121 AP I, art. 48; Nuclear Weapons Advisory Opinion, *supra* note 39, ¶¶ 78–79; Jean-Marie Henckaerts & Louise Doswald-Beck, Customary International Humanitarian Law r. 1, at 3 (Cambridge Univ. Press 2005); *but see* John B. Bellinger III & William J. Haynes II, *A US Government Response to the International Committee of the Red Cross Study Customary International Humanitarian Law*, 89 Int'l Rev. Red Cross 443, 445–47 (2007) (contesting the Study's methodology and NIAC extension of certain Rules), *reprinted in* 46 I.L.M. 514 (2007).

122 Jean-Marie Henckaerts & Louise Doswald-Beck, Customary International Humanitarian Law r. 1 (Cambridge Univ. Press 2005), https://ihl-databases.icrc.org/en/customary-ihl/v1/rule1 (last visited Apr. 16, 2026); *id.* r. 7, https://ihl-databases.icrc.org/en/customary-ihl/v1/rule7 (last visited Apr. 16, 2026).

123 *Id.* rr. 1, 7 (applicability in IAC and NIAC).

international humanitarian law," corresponding to the prohibition in Article 51(4)(c) AP I. Autonomous cyber weapons exhibiting unbounded network propagation through scale–free architectures, where effects cannot be spatially or temporally limited once deployed, fall within Rule 12(c)'s definition and thereby within Rule 11's prohibition as means whose effects cannot be limited as required by IHL.[124]

The central difficulty in cyber operations is that much of the relevant infrastructure, including internet backbone routers, telecommunications switches, commercial data centers, cloud computing platforms, and undersea fiber–optic cables, simultaneously serves civilian and military functions, making the classification exercise under Article 52(2) AP I structurally harder than in most kinetic contexts. Unlike kinetic dual–use objects such as power plants or bridges, where military and civilian functions are often spatially co–located but visually distinguishable, cyber infrastructure may serve both purposes through logically separated traffic flows on physically identical hardware. This renders the distinction unobservable through technical means alone. Article 52(2) defines military objectives as "those objects which, by their nature, location, purpose, or use, make an effective contribution to military action and whose total or partial destruction, capture, or neutralization, in the circumstances ruling at the time, offers a definite military advantage."[125] The dual-use character of cyber infrastructure means that even when military use can be confirmed, the civilian function may remain predominant, and the "circumstances ruling at the time" may change between the access, persistence, and activation phases of a temporally disaggregated cyber operation.[126] Where military purpose was present during the access phase, but civilian use

124 Jean-Marie Henckaerts & Louise Doswald-Beck, Customary International Humanitarian Law r. 11, at 37, r. 12(c), at 40 (Cambridge Univ. Press 2005); AP I, art. 51(4)(c).

125 AP I, art. 52(2).

126 AP I, art. 52(2); 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 2021–2024.

predominates at the activation phase, the object's status at the moment of the attack's effect - not at the moment of emplacement - determines legality under Article 52(2). Pre-positioned implants activated months or years after emplacement are structurally designed to operate across precisely this temporal gap, rendering contemporaneous verification of military objective status structurally unavailable at the operative moment of force. The ICRC Commentary recognizes that infrastructure "dedicated to the production of civilian goods may also be used for the benefit of the army," acquiring "a dual function" of value to both civilians and military forces. In such cases, the object qualifies as a military objective under Article 52(2), but the attack must account for 'the military advantage anticipated' against 'the loss of human life which must be expected among the civilian population and the damage which would be caused to civilian objects."[127] Moreover, the military advantage must be "concrete and direct," not merely "potential or indeterminate," and the determination must be made based on "the circumstances ruling at the time" of attack. In the cyber context, this requires assessing real–time network usage patterns, traffic loads, and routing configurations-information that may be unknowable to an attacker operating through pre–positioned implants activated remotely.[128]

Article 52(3) establishes a rebuttable presumption of civilian status in cases of doubt: "In case of doubt whether an object which is normally dedicated to civilian purposes... is being used to make an effective contribution to military action, it shall be presumed not to be so used."[129] This provision resolves doubt in favor of civilian objects and their protection: even in contact areas, it is prohibited to attack an object subject to the presumption unless it is certain that the object

127 AP I, art. 52(2); 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 2022–2023.

128 AP I, art. 57(2)(a)(iii); 1987 ICRC AP Commentary, *supra* note 18, ¶ 2024; *see also* ICRC, *AWS Selected Issues*, *supra* note 15, at 11, 14; ICRC IHL Challenges Report 2024, *supra* note 19, at 33.

129 AP I, art. 52(3).

accommodates enemy combatants or constitutes a military objective, and strict compliance with the precautionary obligations under Article 57 will, in most cases, confirm either the persisting doubt or the required certainty.[130] The general precautionary obligation of constant care in Article 57(1) is likewise reflected in the Tallinn Manual 2.0, whose Rule 115 requires that, during cyber operations, those who plan or decide upon an attack take feasible precautions to spare the civilian population, individual civilians, and civilian objects from the effects of hostilities.[131]

Ian Henderson and Backstrom's systematic analysis of feasible precautions confirms that feasibility must be assessed not only by reference to the technical capability of the attacking force but also through a contextually determined multi–factor standard encompassing intelligence availability, target characterization, and the operational circumstances at the time of the attack, a framework that, when applied to autonomous cyber weapons, requires assessment of the weapons system's own technical capacity to implement precautionary measures as a component of the feasibility determination.[132] This presumption operates in conjunction with the obligation under Article 57(2)(a)(i) AP I those attackers "do everything feasible to verify that the objectives to be attacked are neither civilians nor civilian objects... but are military objectives." Rule 16's verification requirement poses structural challenges for autonomous systems that cannot perform the semantic, functional assessments necessary to distinguish dual–use cyber infrastructure with the requisite certainty. The ICRC Commentary specifies that those ordering attacks "should at least make sure that the information is precise and recent" and that "in case of doubt, additional information must be requested." The Commentary further clarifies that

130 1987 ICRC AP Commentary, *supra* note 18, ¶ 2034.

131 Tallinn Manual 2.0, *supra* note 18, r. 114 cmt. ¶¶ 1–3 (constant care obligation under AP I art. 57(1) applies during cyber operations; requires continuous attention to civilian protection throughout planning and execution).

132 AP I, arts. 57(2)(a)(i)–(ii); Alan Backstrom & Ian Henderson, *New Capabilities in Warfare: An Overview of Contemporary Technological Developments and the Associated Legal and Engineering Issues in Article 36 Weapons Reviews*, 94 Int'l Rev. Red Cross 483, 485, 492–93 (2012).

even in active combat areas, "there is a presumption that civilian buildings located there are not used by the armed forces, and consequently it is prohibited to attack them unless it is certain that they accommodate enemy combatants or military objects." This verification obligation requires affirmative steps to establish military status: "Those ordering or executing the attack must have sufficient information available to take this requirement into account; in case of doubt, the safety of the civilian population, which is the aim of the Protocol, must be taken into consideration." Autonomous or learning–enabled targeting systems generating probabilistic classifications (e.g., "78% confidence: military server") cannot satisfy this standard, as statistical confidence intervals lack the certainty the Commentary requires and cannot trigger the "additional information" inquiry that Article 57 demands when doubt exists.[133]

The U.S. Department of Defense Law of War Manual reinforces these verification obligations in the cyber context, stating that if cyber operations constitute attacks, "the law of war rules on conducting attacks must be applied," including the duty to "take feasible precautions to verify that the objects of attack are military objectives."[134] The Manual clarifies that "feasible" means "practicable or practically possible, taking into account all circumstances ruling at the time, including humanitarian and military considerations," not "everything possible."[135] Applying this standard to the duty to verify that objects of attack are military objectives, the Manual specifies that "what precautions are feasible depends greatly on the context," including the type of attack and

133 1987 ICRC AP Commentary, *supra* note 18, ¶ 2198 ("in case of doubt, additional information must be requested"); *id.* ¶ 2024 ("Those ordering or executing the attack must have sufficient information available to take this requirement into account; in case of doubt, the safety of the civilian population, which is the aim of the Protocol, must be taken into consideration"); *id.* ¶ 2034 ("it is prohibited to attack them unless it is certain that they accommodate enemy combatants or military objects").

134 DoD Law of War Manual, *supra* note 67, § 16.5.1, at 1033 ("If a cyber operation constitutes an attack, then the law of war rules on conducting attacks must be applied to those cyber operations."); *id.* § 5.5.3 (requiring parties to "take feasible precautions to verify" that objects of attack are military objectives).

135 DoD Law of War Manual, *supra* note 67, § 5.2.3.2 ("Feasible precautions are those that are practicable or practically possible, taking into account all circumstances ruling at the time, including humanitarian and military considerations.").

available time - and expressly recognizes that this obligation is of greater relevance in cyber operations than in conventional warfare, precisely because it applies to a broader set of activities than those governed by the attack rules alone.[136] For instance, when troops are in contact with enemy forces during combat operations, they may need to react immediately based on professional judgment, whereas deliberate targeting processes allow for more extensive verification. Critically, the law of war "does not forestall commanders and other decision-makers from making decisions and taking actions at the speed of relevance.[137] For autonomous cyber weapons, algorithmic opacity may render precautions structurally impossible rather than merely difficult, potentially crossing the threshold from difficult compliance to categorical prohibition.[138]

When the Article 52(3) AP I presumption is interpreted in accordance with this chapter's methodology, Article 31(1) VCLT object-and-purpose interpretation, read in light of the ICJ's holding in Legality of the Threat or Use of Nuclear Weapons that weapons structurally incapable of directing effects at specific military objectives are categorically prohibited, the better view is that residual algorithmic uncertainty cannot be allocated to civilians: Article 52(3) AP I expressly directs that uncertainty in favor of civilian protection. Where an autonomous weapon's probabilistic classification creates commander uncertainty about whether an object satisfies the criteria under Article 52(2) AP I, the presumption under Article 52(3) AP I requires treating the object as civilian absent definitive verification.[139]

---

136 *Id.* § 16.5.3 ("The obligation to take feasible precautions may be of greater relevance in cyber operations than other law of war rules because this obligation applies to a broader set of activities than those to which other law of war rules apply.").

137 DoD Law of War Manual, *supra* note 67, § 5.5.3 ("the law of war, including the requirements discussed in this section, does not forestall commanders and other decision-makers from making decisions and taking actions at the speed of relevance, including in high-intensity conflict, based on their good faith assessment of the information that is available to them at the time").

138 Netta Goussac & Magdalena Pacholska, *The Interpretation and Application of International Humanitarian Law in Relation to Lethal Autonomous Weapon Systems* 15, 20 (UNIDIR 2024), https://unidir.org/publication/the-interpretation-and-application-of-ihl-in-relation-to-laws/ (last visited Apr. 16, 2026) [hereinafter Goussac & Pacholska, *IHL and LAWS*].

139 Vienna Convention on the Law of Treaties art. 31, *opened for signature* May 23, 1969, 1155 U.N.T.S. 331; Nuclear Weapons Advisory Opinion, *supra* note 39, ¶ 78; AP I, art. 52(3).

For autonomous cyber weapons, the challenge of distinguishing intensifies because machine-learning classifiers lack the semantic understanding of functional context that human operators use to resolve dual–use ambiguity. A military logistics database and a civilian hospital patient records system may run identical commercial database software, use identical network protocols, and exhibit structurally similar traffic patterns. An autonomous targeting algorithm trained on network features cannot distinguish between these functionally distinct objects without access to higher–order contextual information, the purpose and use elements of Article 52(2) AP I, that current machine learning architectures cannot reliably extract from network traffic alone.[140]

The Tallinn Manual 2.0 reflects the International Group of Experts' assessment of this challenge in Rule 115, recording expert agreement that, in the context of cyber-attacks, the obligation to do everything feasible to verify that objectives are lawful military targets may require gathering intelligence through network mapping or other processes, and that where such information cannot be obtained, the decision–maker may be required to refrain from conducting the attack or to limit its scope to those components whose status as lawful targets can be sufficiently verified.[141] The doctrinal question is whether "particularly difficult" becomes "structurally infeasible" when autonomous systems must make targeting decisions at machine speed without human semantic understanding. Where an autonomous weapon cannot distinguish between a military command server and a civilian emergency services dispatch, not due to temporary information gaps but due to design–inherent limits of feature–based classification, the

140 AP I, arts. 52(2)–(3); 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 2020, 2024, 2034 ("circumstances ruling at the time," "purpose or use" requiring functional judgment; "unless it is certain" standard; "in case of doubt, the safety of the civilian population . . . must be taken into consideration"); Goussac & Pacholska, *IHL and LAWS*, *supra* note 138, at 11–12, 14–15, 20.

141 Tallinn Manual 2.0, *supra* note 18, r. 115 cmt. ¶¶ 1–5 (verification of targets: obligation to do "everything feasible" to verify that the objective is a lawful military target applies equally to cyber attacks and is not discharged where technical opacity prevents reliable pre-attack assessment).

weapon exhibits the indiscriminate character that the ICJ identified as categorically prohibited in *Nuclear Weapons* paragraph 78.[142]

2. Proportionality: Reverberating Harm and Epistemic Closure

The proportionality rule in Article 51(5)(b) AP I prohibits "an attack which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated."[143] The rule is assessed ex ante on the basis of what the attacker knew or "should have known" at the time of decision, with "expected" referring to what is reasonably foreseeable given available intelligence and military experience.[144]

Rule 14 of the ICRC Customary IHL Study codifies the proportionality principle as customary international law binding in both international and non–international armed conflicts: "Launching an attack which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated, is prohibited."[145] For autonomous cyber weapons, Rule 14's proportionality obligation cannot be delegated to algorithmic systems, given that the U.S. Department of Defense acknowledges that "the legal question of whether the expected incidental harm is excessive may be a 'highly open–ended' inquiry, and the answer may be subjective and imprecise."[146]

---

142 Nuclear Weapons Advisory Opinion, *supra* note 39, ¶ 78; Goussac & Pacholska, *IHL and LAWS*, *supra* note 138, at 15, 20 (dual-use cyber infrastructure classification challenges requiring functional assessment autonomous systems cannot perform).

143 AP I, art. 51(5)(b) (the proportionality rule: attacks are prohibited where "the expected" incidental civilian harm "would be excessive in relation to the concrete and direct military advantage anticipated"; the standard is assessed *ex ante* on the basis of what the attacker knew or should reasonably have known at the time of decision, with "expected" incorporating reasonably foreseeable cascade effects-including cross-domain propagation that commanders should have anticipated based on available technical intelligence).

144 *Id.*; 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 2208–2209.

145 Jean-Marie Henckaerts & Louise Doswald-Beck, Customary International Humanitarian Law r. 14 (Cambridge Univ. Press 2005).

146 DoD Law of War Manual, *supra* note 67, § 5.12.3, at 274.

Cyber operations strain traditional IHL frameworks because civilian harm often arises through indirect and delayed pathways — disruption of electricity, water treatment, healthcare delivery, logistics networks, and communications infrastructure-rather than immediate kinetic effects. The principle of proportionality under customary international law requires that those who plan or decide upon an attack "refrain from deciding to launch any attack which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated."[147] ICRC legal advisers have emphasized that this obligation extends to reasonably foreseeable reverberating effects, stating that "commanders are not only obliged to take into account the direct incidental damage that may be expected from an attack, but must also consider the foreseeable reverberating effects of the attack,"[148] which ICRC analysis defines as effects "that are not directly caused by the attack, but are nevertheless a product thereof," including cascading failures in essential civilian services.[149] In cyber operations, where network interdependencies create unique risks of indirect harm, attackers must assess whether malware will spread beyond intended military targets and take all feasible precautions to avoid and minimize incidental civilian harm, including indirect effects caused by the disruption of interconnected systems.[150]

147 AP I, art. 57(2)(a)(iii).

148 Isabel Robinson & Ellen Nohle, *Proportionality and Precautions in Attack: The Reverberating Effects of Using Explosive Weapons in Populated Areas*, 98 Int'l Rev. Red Cross 107, 113 (2016) [hereinafter Robinson & Nohle, *Reverberating Effects*].

149 Tallinn Manual 2.0, *supra* note 18, r. 113 cmt. ¶¶ 3, 6–7 (indirect and cascading effects to be factored into the proportionality calculation; uncertainty as to collateral damage does not reduce the calculation threshold); Ellen Nohle & Isabel Robinson, *War in Cities: The 'Reverberating Effects' of Explosive Weapons*, ICRC Humanitarian L. & Pol'y Blog (Mar. 2, 2017), https://blogs.icrc.org/law-and-policy/2017/03/02/war-in-cities-the-reverberating-effects-of-explosive-weapons/ (last visited Apr. 16, 2026) [hereinafter Nohle & Robinson, *Reverberating Effects*]; *see also* Jean-Marie Henckaerts & Louise Doswald-Beck, Customary International Humanitarian Law r. 14, at 46, 48–49 (Cambridge Univ. Press 2005) (confirming that the proportionality calculus under customary IHL encompasses reverberating and long-term civilian effects that are reasonably foreseeable at the time of attack, irrespective of whether such effects are direct or indirect consequences of the operation); 1987 ICRC AP Commentary, *supra* note 18, ¶ 1979 (stating that the proportionality assessment requires account to be taken of harm extending beyond the immediate target area, including consequential civilian effects, provided such effects are reasonably foreseeable at the time of planning).

150 AP I, art. 57(2)(a)(ii).

Autonomy intensifies proportionality problems in two distinct ways. First, learning–enabled systems optimize proxy objectives (network disruption, persistent access, lateral movement) without the capacity to internalize civilian protection as a juridical constraint. Proportionality is a legal judgment requiring qualitative assessment of civilian value against military advantage — an assessment that cannot be reduced to algorithmic optimization without fundamentally misconceiving the rule's normative character. As the ICRC Commentary emphasizes, proportionality "must above all be a question of common sense and good faith," requiring military commanders to "carefully weigh up the humanitarian and military interests at stake."[151]

Second, autonomy can make harm pathways uncertain not only in magnitude but in kind, because system behavior adapts in response to real–time network conditions, defensive countermeasures, and unmodeled infrastructure interdependencies. Where an autonomous weapon's propagation or targeting logic adapts dynamically to evade detection or maximize disruption, the "expected" incidental harm at the moment of decision may bear little relationship to the actual harm realized during execution.[152] This creates a temporal and epistemic gap, challenging the premise that proportionality can be meaningfully assessed ex ante.

151 AP I, art. 51(5)(b); 1987 ICRC AP Commentary, *supra* note 18, ¶ 2208; Michael N. Schmitt, *Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics*, 4 Harv. Nat'l Sec. J. Features 1, at 18–19 & n.59, 20–21 (2013), https://harvardnsj.org/wp-content/uploads/2013/06/Schmitt-Reply.pdf (last visited May 12, 2026) [hereinafter Schmitt, *Reply to the Critics*].

152 AP I, art. 51(5)(b).

3. The Cascade Problem: Network Topology and Verification Impossibility

The following analysis proceeds from technical premises; the legal conclusions it reaches are stated explicitly at each stage. Modern critical infrastructure networks often exhibit hub–dependent or scale–free topological properties, where a small number of highly connected nodes (hubs) carry disproportionate traffic and connectivity load.[153] Disruption of these hubs can generate system–wide cascade failures that radiate across sectors and national borders through logical rather than geographic pathways.[154] NotPetya illustrates the point: initially deployed against Ukrainian targets, it propagated within hours to Maersk's global logistics operations and Merck's pharmaceutical manufacturing, disrupting vaccine distribution across six continents, effects theoretically foreseeable in kind, hub-dependent networks are known to amplify disruption, but unpredictable in magnitude and pathway to a degree that precluded the *ex-ante* proportionality assessment Article 51(5)(b) AP I requires.

The analysis identifies two analytically distinct legal challenges. First, some cascades are structurally foreseeable but systematically underestimated: network science establishes that scale–free topologies are vulnerable to hub disruption, making cascade propagation foreseeable in principle, yet the magnitude and cross–sector pathways remain highly uncertain.[155] Under Article 51(5)(b), this does not render proportionality assessment impossible, but it shifts the practical burden: the attacker must demonstrate under conservative modeling assumptions that anticipated military advantage is not expected to be trivial relative to reasonably foreseeable cascade–amplified civilian

153 Albert-László Barabási & Réka Albert, *Emergence of Scaling in Random Networks*, 286 Science 509, 509 (1999) [hereinafter Barabási & Albert, *Scaling in Networks*]; Réka Albert, Hawoong Jeong & Albert-László Barabási, *Error and Attack Tolerance of Complex Networks*, 406 Nature 378, 378–79 (2000) [hereinafter Albert, Jeong & Barabási, *Attack Tolerance*].

154 Albert, Jeong & Barabási, *Attack Tolerance*, *supra* note 153, at 380.

155 Albert, Jeong & Barabási, *Attack Tolerance*, *supra* note 153, at 380; Tallinn Manual 2.0, *supra* note 18, r. 113 cmt. ¶ 13.

harm. Where that demonstration cannot be made under conservative modelling assumptions, the precautionary logic of Article 57(2)(a)(iii) AP I, which requires abandoning an attack when its lawfulness remains in doubt, independently imposes a precautionary prohibition, without requiring any analogical extension of Article 52(3) AP I.

The Merck infection pathway illustrates a different legal problem. NotPetya did not remain confined to financial and administrative systems; it propagated into Merck's manufacturing and research infrastructure, disrupting the production and supply of vaccines and other pharmaceuticals.[156] This domain–crossing cascade pattern is not captured by standard targeting models because it depends on opaque interconnections between corporate IT networks and industrial control systems that planners neither map nor systematically model ex ante. When propagation pathways span functionally distinct network domains in ways that no reasonable targeting methodology can anticipate, the commander cannot satisfy the requirement under Article 57(2)(a)(i) AP I to "do everything feasible" to verify that the chosen means and methods of attack are not expected to cause excessive incidental civilian harm.[157] The violation here is not classic disproportionality (harm exceeding anticipated military advantage), but structural unpredictability that renders compliance with distinction and precaution obligations practically impossible.

156 *Increase Vigilance Against Cyberattacks*, 40 Nature Biotechnology 1155, 1155 (2022), https://doi.org/10.1038/s41587-022-01446-4 (last visited Apr. 16, 2026) (documenting NotPetya's propagation into Merck's manufacturing and research infrastructure); Paul Roberts, *NotPetya Infection Left Merck Short of Key HPV Vaccine*, Security Ledger (Oct. 27, 2017), https://securityledger.com/2017/10/notpetya-infection-left-merck-short-key-vaccine-gardasil/ (last visited Apr. 16, 2026) (reporting Merck forced to borrow Gardasil from CDC stockpile due to production shutdown); Patrick Howell O'Neill, *NotPetya Ransomware Cost Merck More Than $310 Million*, CyberScoop (Oct. 27, 2017), https://cyberscoop.com/notpetya-ransomware-cost-merck-310-million/ (last visited Apr. 16, 2026) (reporting $135 million lost sales and $175 million in additional manufacturing and operating costs).
157 AP I, art. 57(2)(a)(i) (establishing that cascade-induced cross-domain effects create structural impossibility of verification: where opaque interconnections between civilian corporate networks and critical industrial control systems generate cascade harms that no reasonable targeting methodology can model-as illustrated by NotPetya's propagation from financial systems into Merck's pharmaceutical manufacturing infrastructure-commanders cannot satisfy the "do everything feasible" verification standard, rendering the operation non-compliant as a structural matter, not merely as a matter of proportionality).

Translated into the language of Article 57(2)(a)(i) AP I, which requires those who plan or decide upon an attack to "do everything feasible to verify that the objectives to be attacked are neither civilians nor civilian objects and are not subject to special protection but are military objectives within the meaning of paragraph 2 of Article 52," inherently unforeseeable cascade pathways make that verification impossible, not merely difficult.[158] Where structurally foreseeable but systematically underestimated cascade effects prevent that verification from being conducted with the requisite confidence, the attacking State cannot satisfy this standard at the operational moment. These two failure modes carry distinct legal consequences. Where cascades are structurally foreseeable but systematically underestimated by commanders (the first category), proportionality analysis remains applicable but demands a heightened standard: the attacking State should bear the burden of demonstrating that the anticipated concrete and direct military advantage substantially exceeds the reasonably foreseeable cascade–amplified civilian harm, rather than placing the burden on the affected State to prove excessiveness after the fact. By contrast, where cascades are inherently unforeseeable (the second category), proportionality assessment itself becomes impossible: when no reasonable planning methodology can anticipate the range of cross–domain effects, compliance with Article 57(2)(a)(i) AP I verification obligations is, as a structural matter, unattainable.[159]

The critical doctrinal move is to distinguish between proportionality assessments that are difficult but feasible under the standard under Article 57(2)(a) AP I and those that are systematically unreliable to a degree requiring heightened precautionary burdens or, in extreme cases,

158 AP I, art. 57(2)(a)(i) (the verification obligation applies to "those who plan or decide upon an attack"-i.e., at the planning phase, not solely at execution; in the context of autonomous cyber weapons, this extends IHL compliance requirements upstream to design, programming, and procurement, because cascade-induced harms that prevent verification at the operational moment must be foreseen and mitigated at the planning phase or the operation cannot lawfully proceed).

159 *See* ICRC Cyber Operations Report, *supra* note 23, at 2 (Executive Summary).

precautionary prohibition. Article 57(2)(a) AP I requires attackers to take precautions that are "feasible," defined as "that which is practicable or practically possible taking into account all circumstances ruling at the time." Where cascade pathways through interdependent networks exhibit substantial uncertainty due to real–time load dependencies, adaptive routing, and defensive countermeasures that are unknowable at planning time, the 'circumstances ruling at the time' may be insufficient to support a confident ex ante proportionality judgment meeting IHL's humanitarian protective standard.

The ICRC Commentary on Article 51(5)(b) AP I emphasizes that proportionality assessments require "value judgments" that cannot be reduced to mathematical formulae and must be interpreted "in the most restrictive sense," with civilian safety given precedence "in case of doubt."[160] Autonomous cyber weapons optimizing discrete metrics lack the holistic evaluative capacity to assess cumulative effects across multiple infrastructure sectors or to apply precautionary interpretation principles requiring "most restrictive" readings favoring civilian protection.[161]

Article 51(5)(b) AP I prohibits attacks that may be expected to cause incidental civilian loss, injury, or damage that would be excessive in relation to the concrete and direct military advantage anticipated. ICRC commentary interprets "concrete and direct" as requiring a substantial and relatively proximate advantage, thereby excluding remote, speculative, or long–term benefits from the proportionality calculus.[162] Traditional proportionality analysis presumes relatively bounded harm zones that are geographically discrete, temporally contained, and

160 1987 ICRC AP Commentary, *supra* note 18, ¶ 1979; *see also* Int'l Comm. of the Red Cross, *The Principle of Proportionality in the Rules Governing the Conduct of Hostilities Under International Humanitarian Law* 8 (2018), https://www.icrc.org/en/document/international-expert-meeting-report-principle-proportionality (last visited Apr. 16, 2026) [hereinafter ICRC, *Proportionality Study*].

161 ICRC, *AWS Selected Issues*, *supra* note 15, at 5–6, 9, 14; ICRC IHL Challenges Report 2024, *supra* note 19, at 52–53.

162 AP I, art. 51(5)(b); 1987 ICRC AP Commentary, *supra* note 18, ¶ 2209.

causally foreseeable. Autonomous cyber weapons operating through interdependent infrastructure systematically violate each presupposition by propagating through hub–dependent network topologies, generating indeterminate harm zones that expand dynamically as cascade failures spread across ostensibly unrelated systems such as logistics, finance, health, and energy. Harm zones from cascading failures can expand orders of magnitude beyond military planners' initial expectations, such that the concrete and direct military advantage of disabling specific communications nodes becomes trivially small relative to the transnational civilian disruption produced by network–scale propagation.[163]

Where hub-dependent network topologies are present, proportionality assessments must be recalibrated to incorporate cascade-sensitive modeling and precautionary discounting of anticipated military advantage. In practice, this requires concrete, quantitative evidence grounded in conservative modeling assumptions that the anticipated military advantage substantially exceeds reasonably anticipated cascade harm, rather than permitting deployment whenever a non-negligible advantage is merely plausible. This structural uncertainty does not render proportionality assessment under Article 51(5)(b) AP I inapplicable; it demands that the assessment proceed on conservative modeling assumptions that account for foreseeable cascade amplification, rather than on the bounded geographic harm models developed for kinetic targeting. Where such conservative modeling cannot exclude the reasonable expectation of disproportionate cascade harm, the precautionary logic of Article 57(2)(a)(iii) AP I requires abandonment of the attack.

163 *See* ICRC Cyber Operations Report, *supra* note 23, at 2, 7; Matteo Crosignani, Marco Macchiavelli & André F. Silva, *The Propagation of Cyberattacks Through Firms' Supply Networks* 1-2 (Fed. Reserve Bank of N.Y., Staff Report No. 937, 2020), https://www.newyorkfed.org/research/staff_reports/sr937 (last visited Apr. 16, 2026) [hereinafter Crosignani, Macchiavelli & Silva, *Supply Networks*].

4. The Interpretability Constraint: Why Technical Advances Cannot Satisfy the ATO

First, interpretability techniques explain correlations, not legal determinations. LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) can identify which input features most influenced a classification decision, for example, that a network node was classified as a military objective because it exhibited traffic patterns correlating with military command infrastructure.[164] But Article 52(2) AP I requires determining whether an object, "by its nature, location, purpose or use," makes an effective contribution to military action and whether its destruction offers a definite military advantage "in the circumstances ruling at the time."[165] No interpretability technique can translate feature-correlation explanations into the legal assessment. Article 52(2) AP I demands: whether this specific object, in this specific context, satisfies IHL's military objective criteria. Explainability shows why the algorithm

classified, not whether the classification satisfies legal standards.

Second, interpretability techniques themselves introduce new failure modes. Rudin *et al.* demonstrate that post-hoc explanation methods such as LIME and SHAP are structurally unreliable for high-stakes decisions: given the Rashomon set of multiple equally-accurate models, any post-hoc explanation represents one of many possible accounts of a prediction, none of which may accurately reflect the model's true reasoning process.[166] Slack *et al.* demonstrate that post-hoc explanation methods can be manipulated to produce misleading accounts of model behavior: a system may be optimized to generate facially compliant

164 Scott M. Lundberg & Su-In Lee, *A Unified Approach to Interpreting Model Predictions*, *in* 30 Advances in Neural Information Processing Systems 4765, 4766 (NIPS 2017), https://arxiv.org/abs/1705.07874; Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin, *"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, 1136 (2016), https://doi.org/10.1145/2939672.2939778.

165 AP I, art. 52(2); 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 2018, 2024.

166 Cynthia Rudin et al., *Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges*, 16 Stat. Surv. 1, 7–9 (2022), https://doi.org/10.1214/21-SS133.

explanations for targeting decisions that structurally violate distinction or proportionality requirements, defeating rather than enabling IHL verification.[167]

Third, interpretability advances do not address temporal disaggregation or cascade unpredictability. Even if an autonomous cyber weapon could explain why it classified a network node as a military objective at the moment of pre-positioning, this does not enable commanders to verify whether the object remains a military objective months later when effects are activated.[168] Similarly, even if the system can explain its direct targeting decision, this does not predict or explain the cascade effects through interconnected infrastructure that determine whether the attack satisfies proportionality requirements. The structural challenges identified in Sections II.1, II.3, and II.6 — temporal misalignment, network topology vulnerability, and adversarial distribution shift — are architectural features of autonomous cyber operations, not remediable through algorithmic interpretability improvements.

The conclusion is therefore robust against technical optimism objections: the ATO's requirements cannot be satisfied by interpretability advances alone because the incompatibility between autonomous cyber weapons and Article 57 stems from operational characteristics, temporal disaggregation, cascade propagation, and adversarial environments, rather than merely algorithmic opacity. While interpretability research may reduce opacity for certain ML applications, it cannot convert autonomous cyber weapons into means of warfare capable of satisfying IHL's precautionary obligations in the operational contexts analyzed here.[169]

---

167 Dylan Slack et al., *Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods*, *in* Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 180, 180 (2020), https://doi.org/10.1145/3375627.3375830.

168 AP I, art. 57(2)(a)(i); *see* the preceding temporal disaggregation analysis *supra*.

169 ICRC, *AWS Selected Issues*, *supra* note 15, at 7 (recommending prohibition on autonomous weapon systems "whose effects cannot be sufficiently understood, predicted and explained"); AP I, arts. 51(5)(b), 57(2)(a)(i)–(ii).

One structural clarification is essential for the analysis that follows in Chapters 4, 5 and 6. The per se prohibition on Tier 3 deployment established in this Chapter is architecturally specific rather than categorically universal: it prohibits all systems exhibiting all three structural characteristics identified below but does not prohibit machine-assisted or semi-autonomous targeting retaining a human verification capacity at the target-selection node. The prohibition is absolute with respect to systems combining all three features simultaneously: no system that exhibits all three structural characteristics- machine-determined target selection without real-time human verification capacity, behavioral opacity to the deploying commander at the moment of engagement, and technically irreversible propagation effects, can satisfy Article 57(2)(a)(i) AP I as a matter of *lex lata.* The design architecture proposed in Chapter 5 — kill-switches, geofencing, temporal deactivation logic, and mandatory human verification loops for specially protected objects - is not a set of modifications that render a Tier 3 system IHL-compliant under existing law. It is the set of design requirements that, if satisfied, would reclassify a putative Tier 3 system as Tier 2 by eliminating the specific architectural features, behavioral opacity, irreversible propagation, and the absence of real-time human verification capacity that produce the structural incompatibility. Chapter 5's Human Security by Design framework is accordingly the legal pathway from prohibited Tier 3 to permissible Tier 2: it specifies the architectural conditions under which an autonomous system crosses the threshold above which the precautionary framework of AP I can again operate as a legal constraint. The per se prohibition is directed at systems as currently configured; the design framework addresses how such systems must be reconfigured before they may be lawfully authorized. The distinction between these two categories, the per se prohibited Tier 3 system and the conditionally permissible HSD-architecture system, corresponds to the distinction between per

se prohibition and conditional permissibility that the proposed IAWCO framework makes operational in Chapter 6. Whether this precautionary allocation of doubt constitutes a present treaty obligation under Article 51(5)(b) AP I, read in light of the *in dubio pro civili* principle and the object-and-purpose standard of Article 31(1) VCLT, or represents a permissible *de lege interpretanda* development of the proportionality standard, is addressed in the normative-tier framework at the opening of this Chapter. This dissertation proceeds on the *lex lata* conclusion that, where cascade pathways cannot be reliably bounded *ex ante*, Article 57(2)(a)(i) verification is structurally infeasible, and the authorization of deployment constitutes a breach of the precautionary obligation at that moment.

5. Military Necessity as a Putative Exception: Why It Does Not Resolve the Structural Incompatibility

A State might argue that military necessity — the background principle of IHL permitting measures necessary for achieving a legitimate military objective and not otherwise prohibited - justifies the deployment of Tier 3 ACW in circumstances where no alternative means can achieve the military objective within operationally acceptable timeframes. On this argument, the speed and scale advantages of autonomous cyber weapons are not merely operationally convenient but militarily necessary in the sense that manual human authorization cycles cannot respond to machine-speed threat environments without operationally unacceptable delay. This argument deserves direct address because, if accepted, it would constitute the most significant doctrinal challenge to the structural incompatibility thesis. The argument must be rejected because it conflates two distinct functions of military necessity in IHL doctrine. Military necessity operates, first, as a background principle defining the outer limit of IHL's tolerance for collateral harm — it informs the proportionality calculus and is built into the "military advantage" standard of Article

51(5)(b) AP I. It operates, second, in the argument now under consideration, as a putative derogation clause capable of overriding specific treaty rules whose observance would be operationally inconvenient. The prohibition admits no such exception: Article 57(2)(a)(i) AP I contains no derogation clause, and the ICRC Commentary confirms both that the "everything feasible" standard is not qualified by operational convenience and that "there is no implicit clause in the Conventions which would give priority to military requirements." The rule that commanders must "do everything feasible to verify" target status is accordingly not qualified by any exception for operationally inconvenient circumstances. As the ICJ stated in the Nuclear Weapons Advisory Opinion, the precautionary obligations of Article 57 derive from and give operational effect to the cardinal principle of distinction, which the Court identified as among the "intransgressible principles of international customary law" that bind all States without exception.[170] The structural incompatibility thesis does not claim that Tier 3 deployment violates Article 57(2)(a)(i) AP I because human verification would be operationally inconvenient or costly. It claims something categorically different: that the architectural design of Tier 3 systems eliminates the factual conditions under which the Article 57(2)(a)(i) obligation can operate as a legal duty at all. Military necessity speaks to what belligerents may do within the framework of IHL's precautionary architecture; it does not speak to whether that architecture's preconditions for operation are present. Where a weapon's design eliminates real-time human verification capacity as a matter of

170 AP I, art. 57(2)(a)(i) (containing no derogation clause qualifying the "everything feasible" verification obligation); 1987 ICRC AP Commentary, *supra* note 18, ¶ 2198 ("everything that was practicable or practically possible, taking into account all the circumstances at the time of the attack, including those relevant to the success of military operations"); *id.* ¶ 2206 ("There is no implicit clause in the Conventions which would give priority to military requirements. The principles of the Conventions are precisely aimed at determining where the limits lie; the principle of proportionality contributes to this."); Nuclear Weapons Advisory Opinion, *supra* note 39, ¶¶ 78–79 ("The cardinal principles contained in the texts constituting the fabric of humanitarian law" "constitute intransgressible principles of international customary law"). The characterization in ¶ 79 refers directly to the cardinal principles of distinction and the prohibition of indiscriminate weapons; its application to the precautionary obligations of AP I art. 57 rests on the derivative relationship between art. 57's verification requirement and the cardinal principle of distinction, art. 57 constituting the operational mechanism through which distinction is implemented at the targeting stage.

architecture, the question is not whether military necessity permits the commander to dispense with verification; the question is whether there is a legally cognizable human verifier present at the moment of engagement at all. There is not. Military necessity cannot supply the human judgment that IHL presupposes; it can only affect what that judgment is permitted to conclude.

6. Precaution under Article 57 AP I: Machine–Learning Failure Modes and Verification Impossibility

Article 57 AP I operationalizes precaution through the four obligations established above: target verification under Article 57(2)(a)(i), means and methods selection under Article 57(2)(a)(ii), cancellation and suspension under Article 57(2)(b), and advance warning under Article 57(2)(c).[171] For autonomous cyber weapons, the machine-learning failure modes analyzed below render each of these obligations structurally impossible to satisfy.

The legal standard governing these four obligations is one of due diligence and good faith, not of absolute certainty: no guarantee of perfect identification can be demanded, but the law requires that those who plan or decide upon an attack take all practicable measures to verify target status at the moment of each specific engagement.[172]

Due diligence constitutes a chapeau obligation over the full precautionary architecture of Article 57 AP I as applied to autonomous cyber capabilities. It requires States to undertake all appropriate measures at the research, development, programming, and deployment stages, not merely at the moment of operational execution. Operational

171 AP I, arts. 57(2)(a)(i), 57(2)(a)(ii), 57(2)(b), 57(2)(c); 1987 ICRC AP Commentary, *supra* note 18, ¶ 2195 ("in case of doubt, even if there is only slight doubt, they must call for additional information and if need be give orders for further reconnaissance"); *id.* ¶¶ 2200–2203 (means and methods selection); *id.* ¶¶ 2220–2221 (cancellation and suspension); *id.* ¶¶ 2222–2225 (advance warning). Article 57(2)(c) alone contains the qualification "unless circumstances do not permit," adopted by majority over proposals for stronger obligation language-a structural distinction absent from articles 57(2)(a)(i) and (b); the analysis of each obligation below proceeds in accordance with that obligation's individual textual structure.

172 AP I, art. 57(2)(a)(i) ("do everything feasible to verify").

constraints embedded ex ante in system design, deactivation thresholds, fail-safe override mechanisms, and target-profile limits, constitute the primary diligent measures available where real-time human control is architecturally absent.[173]

For autonomous cyber weapons, these four obligations create cumulative verification challenges that machine learning failure modes render structurally impossible to satisfy. Distribution shift causes targeting accuracy to degrade under operational conditions, while confidence metrics mask this degradation, undermining commanders' ability to verify the status of military objectives under Article 57(2)(a)(i) AP I. Adversarial manipulation enables hostile actors to induce civilian targeting errors through imperceptible input perturbations, rendering means/methods selection under Article 57(2)(a)(ii) AP I unreliable when adversaries can exploit algorithmic vulnerabilities faster than human operators can detect misclassifications. Specification gaming causes autonomous systems to optimize proxy objectives diverging from intended goals, generating civilian harm that commanders cannot anticipate ex ante and therefore cannot cancel or suspend under Article 57(2)(b) AP I. Network propagation at machine speed precludes effective advance warning under Article 57(2)(c) AP I when cyber operations traverse global infrastructure faster than notification mechanisms can function.

These challenges are not remediable through improved training data or refined algorithms. They arise from fundamental characteristics of machine learning systems operating in adversarial environments where distribution shift, adversarial manipulation, and specification gaming interact multiplicatively to render Article 57's precautionary framework

173 Marta Stroppa, *The Use of Autonomous Cyber Capabilities in Armed Conflict and States' Due Diligence Obligations to Avoid Unintended Violations of International Humanitarian Law*, Int'l Rev. Red Cross, 2026, at 4–5 (chapeau obligation), 5 (development and deployment stages), 21 (operational constraints as diligent measures), https://doi.org/10.1017/S181638312610109X).

structurally inoperable. The following analysis examines each failure mode and its legal consequences for IHL compliance.

Machine learning systems optimize performance on training datasets representing historical adversary systems and infrastructure configurations. Upon deployment, autonomous cyber weapons encounter novel target environments exhibiting different feature distributions, a phenomenon known as distribution shift. Hendrycks and Dietterich's robustness benchmarking demonstrated that State–of–the–art classifiers experience substantial performance degradation under corruptions resembling operational perturbations, with robustness to corruption improving only slightly from AlexNet to ResNet, leaving current classifiers "decidedly subhuman" despite superhuman performance on clean training data.[174] For autonomous cyber weapons, the distribution shift creates serious challenges to compliance with Article 57(2)(a) of AP I: misclassification rates increase in operational environments while system outputs remain unchanged, rendering accurate post–deployment verification of IHL compliance impossible; and because distribution magnitude cannot be reliably predicted in advance, autonomous cyber weapons create Article 36 AP I verification deficits that preclude the legal certainty required before deployment.[175] Even under Quéguiner's authoritative interpretation requiring commanders to obtain "the most reliable information possible," commanders cannot satisfy this standard when algorithmic systems exhibit calibration failures that render confidence metrics unreliable under distribution shift.[176] The ICRC's institutional

174 Dan Hendrycks & Thomas G. Dietterich, *Benchmarking Neural Network Robustness to Common Corruptions and Perturbations*, *in* Proceedings of the International Conference on Learning Representations 1, 6 (2019), https://openreview.net/pdf?id=HJz6tiCqYm (last visited Apr. 16, 2026) [hereinafter Hendrycks & Dietterich, *Benchmarking*].

175 AP I, arts. 36, 57(2)(a)(i); Hendrycks & Dietterich, *Benchmarking*, *supra* note 174, at 1, 6; ICRC, *AWS Selected Issues*, *supra* note 15, at 7, 17–18.

176 Jean-François Quéguiner, *Precautions Under the Law Governing the Conduct of Hostilities*, 88 Int'l Rev. Red Cross 793, 797–98 (2006); Yaniv Ovadia et al., *Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift*, 33 Advances in Neural Information Processing Systems 14003, at 1, 5–6, 9 (NeurIPS 2019), https://proceedings.neurips.cc/paper_files/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf (last visited Apr. 16, 2026) [hereinafter Ovadia et al., *Predictive Uncertainty*].

analysis confirms that autonomous systems must "predictably and reliably operate as intended" and that "the greater the uncertainty and unpredictability, the greater the risk that IHL might be violated."[177] The multilateral record synthesized by the UNIDIR background paper on IHL and LAWS reinforces this conclusion, recording broad acceptance among States and experts that "in order to comply with the targeting rules, including the duty to prevent breaches, commanders and users need to have an appropriate level of understanding of the system's functioning.[178] However, empirical research demonstrates that distribution shift causes neural network accuracy to degrade substantially under operational perturbations, with architectural improvements yielding "negligible changes in relative corruption robustness," leaving classifiers performing substantially below the human baseline under operational perturbation conditions, a degradation masked rather than signaled by the system's own outputs.[179] Ovadia et al.'s independent empirical confirmation that predictive uncertainty estimates become unreliable under dataset shift reinforces this conclusion: commanders cannot rely on confidence metrics validated during development when systems encounter novel operational conditions.[180] Adversaries deliberately exploit machine learning system opacity through adversarial manipulation, the craft of generating imperceptible input perturbations that cause targeted misclassifications. Goodfellow, Shlens, and Szegedy's foundational demonstration that small, imperceptible perturbations cause State-of-the-art classifiers to misidentify inputs with high confidence establishes that adversarial vulnerability is not an incidental defect but a structural feature of

177 Neil Davison, *Autonomous Weapon Systems Under International Humanitarian Law*, UNODA Occasional Papers No. 30, at 10, 14–15 (Nov. 2017), https://www.icrc.org/sites/default/files/document/file_list/autonomous_weapon_systems_under_international_humanitarian_law.pdf (last visited Apr. 16, 2026) [hereinafter Davison, *AWS Under IHL*].

178 Goussac & Pacholska, *IHL and LAWS*, *supra* note 138, § 5.3, at 21 ("From an IHL perspective, it seems widely accepted that, in order to comply with the targeting rules, including the duty to prevent breaches, commanders and users need to have an appropriate level of understanding of the system's functioning").

179 Hendrycks & Dietterich, *Benchmarking*, *supra* note 174, at 1, 6.

180 Ovadia et al., *Predictive Uncertainty*, *supra* note 176, at 1, 5, 9.

gradient-based learning architectures.[181] For autonomous cyber weapons, adversarial manipulation means a hostile actor can craft network traffic patterns causing an autonomous targeting system to misclassify a civilian hospital server as a military command node, or classify a military objective as civilian, with the system reporting high confidence in either direction. This renders the means-selection obligation under Article 57(2)(a)(ii) AP I structurally unreliable in adversarial operational environments: the commander cannot select means that avoid civilian harm when the means themselves can be weaponized by the adversary to generate precisely that harm.

While Goodfellow *et al.* demonstrated that imperceptible perturbations cause high-confidence misclassification, Nicholas Carlini & David Wagner subsequently demonstrated that targeted misclassifications - causing the system to output a specific adversary-chosen incorrect class - can be reliably produced even against defences designed to prevent such attacks, confirming that adversarial manipulation is not incidentally possible but systematically achievable against deployed systems.[182] This creates violations of Article 48 AP I-based distinctions when adversarial inputs induce civilian-targeting errors. Translating this attack vector to autonomous cyber weapons: An adversary can craft network packets, malware artifacts, or infrastructure metadata that appear legitimate to human analysts but cause an autonomous cyber weapon's targeting classifier to misclassify civilian infrastructure as military objectives. The adversary need not understand the internal mechanics of the machine learning model; black–box adversarial example generation techniques work even without access to the model's

181 Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy, *Explaining and Harnessing Adversarial Examples*, *in* Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), arXiv:1412.6572v3, at 1–2 (Mar. 20, 2015), https://arxiv.org/abs/1412.6572.

182 Nicholas Carlini & David Wagner, *Towards Evaluating the Robustness of Neural Networks*, *in* 2017 IEEE Symposium on Security and Privacy 39, 39–40, https://arxiv.org/abs/1608.04644 (last visited Apr. 16, 2026) [hereinafter Carlini & Wagner, *Robustness* (preprint)] ("we demonstrate that defensive distillation does not significantly increase the robustness of neural networks by introducing three new attack algorithms that are successful on both distilled and undistilled neural networks with 100% probability").

architecture or parameters. An adversary observes the autonomous system's outputs to a series of test inputs, uses these outputs to infer the underlying decision boundary, and generates adversarial examples that cross that boundary without human–perceptible changes to the underlying data.[183] This adversarial vulnerability creates three challenges to IHL compliance: intentional evasion of civilian protection becomes possible through adversarial inputs designed to induce targeting errors; ex ante verification at Article 36 review cannot assess weapons that are vulnerable to adversarial exploitation in operational environments, and Article 57 verification becomes impracticable in high–speed cyber operations, where adversaries can manipulate targeting inputs faster than human operators can detect and correct misclassifications. Krakovna *et al.* systematically documented specification gaming, the failure mode in which machine learning systems optimize proxy objectives that diverge from intended goals, demonstrating that this phenomenon arises across a wide range of deployed reinforcement learning systems, not merely in laboratory settings.[184] The deployment of the DoD's Algorithmic Warfare Cross–Functional Team (Project Maven) exemplifies this failure mode in a live military AI targeting context: the programme's stated criterion for fielding algorithmic systems was "to get the best possible algorithmic systems into the battlefield quickly, even when they were just 80% complete,"[185] operating by scanning drone footage for pattern–

183 Nicholas Carlini & David Wagner, *Towards Evaluating the Robustness of Neural Networks*, *in* 2017 IEEE Symposium on Security and Privacy 39, 39–40, https://doi.org/10.1109/SP.2017.49 (last visited Apr. 16, 2026) [hereinafter Carlini & Wagner, *Robustness* (published)].

184 Victoria Krakovna et al., *Avoiding Side Effects in Complex Environments*, 34 Advances in Neural Information Processing Systems 1 (NeurIPS 2020) (demonstrating that specification gaming-the optimization of proxy objectives diverging from intended goals-arises systematically across reinforcement learning systems, not only in laboratory conditions); *see also* Dario Amodei et al., *Concrete Problems in AI Safety*, arXiv:1606.06565, at 3–5 (working paper 2016) (identifying reward hacking and goal misspecification as related failure modes).

185 Kate Crawford, Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence 189 (Yale Univ. Press 2021) ("The goal of the program was to get the best possible algorithmic systems into the battlefield quickly, even when they were just 80% complete") [hereinafter Crawford, *Atlas of AI*]. The passage describes the operational fielding criterion applied to Project Maven; the formulation has not been verified against a primary government document or Congressional testimony transcript as of the date of submission and is cited through secondary sources. If a primary-source document becomes available, this footnote should be updated accordingly.

correlated categories rather than performing the legal determination that a given object constitutes a military objective under Article 52(2) AP I, substituting a proxy metric (pattern confidence) for the IHL–mandated targeting judgment that Article 57(2)(a)(i) AP I requires.[186] Such specification gaming produces civilian harm orders of magnitude exceeding anticipated military advantage, violating the proportionality threshold under Article 51(5)(b) AP I.

These three failure modes, adversarial manipulation, distribution shift, and specification gaming, interact multiplicatively, each independently defeating a distinct element of the ATO's three-component verification structure, to create a cumulative verification crisis rendering autonomous cyber weapons structurally incompatible with IHL's protective architecture. These modes collectively render Article 57(2)(a) AP I, requiring commanders to 'do everything feasible to verify' target classifications, practically impossible as an operational matter. Verification requires access to complete information about: (i) training data distributions and their relationship to operational distributions; (ii) adversarial robustness properties and potential adversarial attack vectors; (iii) objective function specifications and their alignment with underlying military objectives; and (iv) specification gaming vulnerabilities. This information remains proprietary (known only to system developers and contractors), partially unknown even to developers (because machine learning system behavior emerges from complex training interactions), and fundamentally incomplete (because adversarial examples evolve after deployment through operational interaction with novel adversaries).[187]

---

186 AP I, arts. 52(2), 57(2)(a)(i); *see* 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 2018–2024 (explaining that art. 57(2)(a)(i) requires those who plan or decide upon attacks to verify that objectives are military objectives within the meaning of art. 52(2) before attack is executed).

187 AP I, art. 57(2)(a)(i); ICRC, *AWS Selected Issues*, *supra* note 15, at 4–5, 7, 17; Jonathan Kwik, *Iterative Assessment for Military Artificial Intelligence (AI) Systems* 8, 11–12 (Asser Inst. Research Paper 2025-03, 2025), https://www.asser.nl/media/798036/5338856.pdf (last visited Apr. 16, 2026) [hereinafter Kwik, *Iterative Assessment*]; Dario Amodei et al., *Concrete Problems in AI Safety*, arXiv:1606.06565, at 2 (working paper 2016), https://arxiv.org/abs/1606.06565.; Ovadia et al., *Predictive Uncertainty*, *supra* note 332, at 2, 5, 7, 9.

The legal consequence is as follows: If, despite taking all feasible precautions under Article 57(2)(a) AP I, a military commander cannot sufficiently understand, predict, or explain a system's targeting effects before deployment and operational execution, the system's use violates IHL compliance obligations. Critically, design–inherent opacity (where ML architecture precludes interpretability by engineering design, not merely by current technological limitations) renders such systems indiscriminate by nature. The ICRC specifically holds that "an AWS that precludes a human user from being able to understand, predict or explain its output cannot be used in compliance with IHL because its design, performance or operating features would render its use tantamount to blind firing... making it indiscriminate by nature."[188]

The implications of this structural opacity for command responsibility under Article 28(a) Rome Statute, including the deployment-stage reorientation of the accountability inquiry, are developed in Chapter 4, Section I.3. The ICRC's finding that an autonomous weapon system "that precludes a human user from being able to understand, predict or explain its output cannot be used in compliance with IHL because its design, performance or operating features would render its use tantamount to blind firing"[189] is, on this reading, not merely a technical compliance observation: it is a predicate for command responsibility at the authorization stage. The ATO's ex ante interpretability requirement therefore does not impose a new obligation beyond existing IHL; it defines what the existing command responsibility doctrine already requires of commanders who authorize opaque autonomous systems.

188 ICRC, *AWS Selected Issues*, *supra* note 15, at 7.

189 Int'l Comm. of the Red Cross, *Autonomous Weapon Systems and International Humanitarian Law: Selected Issues* 10 (Mar. 2026), https://www.icrc.org/sites/default/files/2026-03/4896_002_Autonomous_Weapons_Systems_-_IHL-ICRC.pdf (last visited Apr. 16, 2026) [hereinafter ICRC, AWS and IHL (2026)] ("An AWS that precludes a human user from being able to understand, predict or explain its output cannot be used in compliance with IHL because its design, performance or operating features would render its use tantamount to blind firing".

This indiscriminacy becomes permanent and hence prohibitive under Article 36 weapons review, distinguishing design–inherent unpredictability from temporary opacity remediable through interpretability research. The distinction matters legally: A system exhibiting emergent behaviors where "it is difficult to accept that a commander could reasonably anticipate the effects" the system would have is likely indiscriminate by nature.[190]

The ICJ held that "States must never... use weapons that are incapable of distinguishing between civilian and military targets."[191] Whether this prohibition articulated in the context of weapons with inherently indiscriminate physical effects extends to weapons exhibiting algorithmic opacity remains contested among scholars.[192] This indiscriminacy–by–opacity thesis finds support in authoritative IHL scholarship. Michael Schmitt, while generally skeptical of categorical autonomous weapons prohibitions, acknowledges that "if a weapon system cannot reliably distinguish between military objectives and civilian objects or persons in the circumstances ruling at the time of the attack, it may not be used because it is indiscriminate by nature."[193] William Boothby articulates the verification challenge for autonomous and automated weapon systems by insisting that a weapon may not be procured or employed where it cannot be used in circumstances that allow the precautionary obligations in Article 57(2)(a)(i)–(ii) AP I to be fulfilled, and that an autonomous or automated weapon system which cannot make the determinations required by those provisions "fails to

190 ICRC, *AWS Selected Issues*, *supra* note 15, at 7; *see also* Kwik, *Iterative Assessment*, *supra* note 187, at 12, 21.

191 Nuclear Weapons Advisory Opinion, *supra* note 39, ¶ 78 (ICJ's formulation of the indiscriminate weapons prohibition: "States must never . . . use weapons that are incapable of distinguishing between civilian and military targets"; this formulation constitutes the foundational authority for the indiscriminacy-by-opacity thesis-where algorithmic opacity renders an autonomous cyber weapon structurally incapable of making the civilian/military distinction required by AP I arts. 51(4)(b)–(c), the prohibition attaches regardless of the technology type or kinetic/non-kinetic character of the effect produced).

192 Int'l Comm. of the Red Cross, *ICRC Position on Autonomous Weapon Systems* 3, 7 (May 2021), https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems (last visited Apr. 16, 2026) [hereinafter ICRC, *2021 AWS Position*]; *compare* Schmitt, *Reply to the Critics*, *supra* note 151, at 10 *with* Rebecca Crootof, *Autonomous Weapon Systems and the Limits of Analogy*, 9 Harv. Nat'l Sec. J. 51, 57 (2018) [hereinafter Crootof, *Limits of Analogy*].

193 Schmitt, *Reply to the Critics*, *supra* note 151, at 10–11.

meet the precautions obligations and should either not be procured or at least should not be employed in the relevant circumstances."[194] Applied to autonomous cyber weapons exhibiting distribution shift, adversarial manipulation, and specification gaming, these doctrinal positions converge with the ICRC's assessment that an AWS which precludes a human user from understanding, predicting, or explaining its output is tantamount to blind firing and therefore indiscriminate by nature, so that where commanders cannot predict, explain, or verify targeting effects, the weapon cannot satisfy the discrimination and precautions rules and is prohibited under Article 51(4)(b)–(c) and Article 57(2)(a) AP I.[195] This doctrinal convergence extends beyond IHL scholarship to encompass the ethics and law of emerging military technologies: George R. Lucas Jr. demonstrates that autonomous weapons capable of AI–augmented operation "will not be able to abide by the laws of armed conflict, because they will not be able to meet basic requirements of LOAC, such as the principles of distinction, proportionality, and precaution," and that such systems "will likely further distort or short–circuit any system of accountability for violations of LOAC, creating a 'responsibility gap'," a structural diagnosis that applies with particular force to autonomous cyber weapons whose opacity defeats both prospective IHL verification and retrospective accountability attribution.[196]

194 William H. Boothby, Weapons and the Law of Armed Conflict 254–55 (2d ed., Oxford Univ. Press 2016) ("If, therefore, one of these obligations can be fulfilled using a manned mission but cannot be fulfilled using an autonomous mission, the manned mission should be employed, and the weapons reviewer should draw attention to this requirement in the weapon review" and concluding that "an autonomous or automated weapon system that is not capable of making that determination fails to meet the precautions obligations and should either not be procured or at least should not be employed in the relevant circumstances"); *see also id.* at 252–54 (requiring the reviewer to determine whether the system is "capable of being used in accordance with the targeting rules" and, in particular, with the precautions required in AP I art. 57).

195 AP I, arts. 51(4)(b)–(c), 57(1), 57(2)(a); ICRC, *AWS and IHL (2026)*, *supra* note 189, at 9–10 (stating that "an AWS that precludes a human user from being able to understand, predict or explain its output cannot be used in compliance with IHL because its design, performance or operating features would render its use tantamount to blind firing" and that machine learning-based AWS whose functioning can change after the commencement of an attack in ways unforeseen to the human user "would likely be indiscriminate by nature").

196 George R. Lucas, Law, Ethics and Emerging Military Technologies: Confronting Disruptive Innovation 91 (Routledge 2023) [hereinafter Lucas, *Law, Ethics and Emerging Military Technologies*].

This chapter contends that the debate conflates two distinct issues: whether algorithmic opacity alone renders a weapon indiscriminate under Article 51(4) AP I, and whether algorithmic opacity combined with operational unpredictability violates Article 57's verification requirements. While the former remains contested, the latter is doctrinally settled: a weapon whose targeting effects cannot be verified at the moment of attack fails Article 57(2)(a)'s "everything feasible" standard, rendering its use unlawful regardless of whether the weapon itself is classified as inherently indiscriminate.[197] This formulation establishes that a weapon's inherent incapacity for distinction, not merely its occasional misuse, constitutes a legal barrier to deployment.[198] Where autonomous cyber weapons exhibit design–inherent opacity precluding commanders from understanding, predicting, or explaining targeting effects, such systems are indiscriminate by nature under the ICJ's formulation. They cannot reliably distinguish between civilian and military targets because their decision–making process is opaque. This indiscriminacy is permanent (inherent in the system design, not remediable through training improvements) and hence prohibitive under Article 36 AP I weapons review, triggering categorical prohibition under Article 51(4)(b)–(c) AP I.[199]

197 AP I, arts. 51(4)(b)–(c), 57(2)(a)(i); 1987 ICRC AP Commentary, *supra* note 18, ¶ 2198 (interpreting "everything feasible" to require "practicable or practically possible" measures and "necessary identification measures in good time"); ICRC, 2021 AWS Position, *supra* note 192, at 7 (stating AWS that "preclude a human user from being able to understand, predict or explain its output cannot be used in compliance with IHL").

198 Nuclear Weapons Advisory Opinion, *supra* note 39, ¶ 78 (applying the ICJ's formulation to the inherent-incapacity threshold: the prohibition attaches where a weapon's design renders it constitutionally incapable of distinction-as opposed to a weapon capable of distinction but misused in a particular instance; autonomous cyber weapons exhibiting permanent algorithmic opacity satisfy this threshold because their targeting process is design-inherently, not contingently, incapable of the civilian/military distinction that AP I arts. 51(4)(b)–(c) requires, and no operational procedure, training improvement, or human oversight protocol can remediate a design-inherent incapacity).

199 AP I, arts. 51(4)(b)–(c); ICRC, *AWS Selected Issues*, *supra* note 15, at 10, 17–18.

7. The Doctrinal Foundations of the ATO

The ATO: Doctrinal Content. Article 57(2)(a)(i) of AP I requires those who plan or decide upon attacks to "do everything feasible to verify that the objectives to be attacked are neither civilians nor civilian objects and are not subject to special protection but are military objectives within the meaning of paragraph 2 of Article 52.[200] Article 52(2) in turn limits military objectives, "in so far as objects are concerned," to those which by their nature, location, purpose or use make an effective contribution to military action and whose destruction, capture or neutralization, "in the circumstances ruling at the time," offers a definite military advantage.[201] The ICRC Commentary's authoritative interpretation confirms that "everything feasible" denotes "everything practicable or practically possible, taking into account all the circumstances at the time of the attack," deliberately replacing the earlier and less demanding formulation of "reasonable steps"; commanders must act on "all circumstances ruling at the time" and must obtain "the most reliable information possible" before authorizing an engagement.[202] The ICRC's 2026 position paper on autonomous weapon systems and IHL translates these requirements into the AWS context with particular analytical clarity: because IHL obligations regarding the conduct of hostilities must be fulfilled by human commanders and combatants who must determine the lawfulness of the attacks they plan, decide upon or execute and remain accountable for these determinations, an AWS that precludes the user from knowing which specific person or object will trigger a strike, precisely when and where the strike will occur, or who and what will be affected renders compliance with Article 57(2)(a)(i) AP I structurally

200 AP I, art. 57(2)(a)(i) (treaty text establishing the verification obligation that generates the algorithmic transparency duty: attackers must "do everything feasible to verify" that targets qualify as military objectives within art. 52(2); where an autonomous cyber weapon's decision-making process is opaque to commanders, this obligation cannot be satisfied because commanders cannot confirm that the system's classification output corresponds to the art. 52(2) legal standard-creating the Cognitive-agency standard as a necessary, not merely desirable, condition of lawful autonomous targeting).

201 AP I, art. 52(2); Jean-Marie Henckaerts & Louise Doswald-Beck, Customary International Humanitarian Law r. 8 (Cambridge Univ. Press 2005).

202 AP I, art. 57(2)(a)(i); 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 2195, 2198.

impossible."[203] Some States and commentators have cautioned that this formulation risks rigidifying the flexible "everything feasible" standard into a categorical epistemic threshold; this dissertation proceeds on the view that, in light of the treaty text and the Commentary's explicit rejection of a mere reasonableness standard, the ICRC formulation articulates rather than extends the content of Article 57's precautionary obligations when applied to systems exhibiting structural opacity.[204]

From these treaty–based and authoritative institutional foundations, the ATO may be extracted and articulated as a three–component obligation, an original doctrinal contribution derived from Article 52(2), 57(2)(a)(i) and 57(2)(b) AP I, and not yet named in this form in the existing literature. First, the ex–ante interpretability component requires that, before deployment, those who plan or decide upon attacks must possess sufficient technical understanding of the system's classification logic, training data constraints, and operational parameters to perform the full verification required by Article 57(2)(a)(i), that is, to determine that everything feasible has been done to confirm that the prospective target is not a civilian object or an object under special protection, but constitutes a military objective within the meaning of Article 52(2) "in the circumstances ruling at the time" under foreseeable operational conditions. Second, the in–operandi auditability component requires that, during operations, the system's targeting decisions remain sufficiently transparent to permit real–time assessment of whether engagement thresholds remain consistent with the distinction and proportionality obligations, thereby enabling the commander to satisfy the cancellation and suspension duty under Article 57(2)(b). Third, the ex post forensic recoverability component, the element whose absence in current State practice is most directly demonstrated by the three

203 Int'l Comm. of the Red Cross, *Autonomous Weapon Systems and International Humanitarian Law* 8 (2026), https://www.icrc.org/en/document/autonomous-weapon-systems-and-international-humanitarian-law; AP I, art. 57(2)(a)(i).

204 AP I, art. 57(2)(a)(i); 1987 ICRC AP Commentary, *supra* note 18, ¶ 2198; Vienna Convention on the Law of Treaties art. 31(1), *opened for signature* May 23, 1969, 1155 U.N.T.S. 331.

operational cases examined in Section II.8 below, requires that the basis for each targeting decision be reconstructed after the fact to enable accountability attribution through the frameworks established by Article 4, 7, 8 and 14(3) of the ARSIWA on State responsibility and by Article 8(2)(b)(i)–(ii), 8(2)(b)(iv) and 17 of the ICC Rome Statute's complementarity regime. Where any of these three components cannot be satisfied by reason of a system's design inherent opacity, the ATO (as a matter of Article 31(1) VCLT systematic interpretation of Article 57(2)(a)(i) AP I, as established in this Section) requires that the weapon not be employed, and the Article 36 weapons review must conclude that the system cannot be lawfully employed.[205]

The three components of the ATO map directly onto the precautionary obligations of Tallinn Manual 2.0 Rules 114 through 119, which articulate the most operationally demanding IHL constraints applicable to cyber-attack operations under customary international law and Article 57 AP I, and which the International Group of Experts concluded, with the degree of consensus noted in each Rule's commentary, reflect customary international law in both international and non-international armed conflicts.[206]

The *ex-ante* interpretability component operationalizes the "everything feasible" target verification standard of Rule 115: it defines, with

205 AP I, arts. 36, 52(2), 57(2)(a)(i), 57(2)(b); ARSIWA, *supra* note 21, arts. 4, 7, 8, 14(3).

206 Tallinn Manual 2.0, *supra* note 18, r. 114 cmt. ¶ 4, at 477 (constant care, based on AP I art. 57(1), confirmed customary in both conflict types; duty "requires commanders and all others involved in the operations to be continuously sensitive to the effects of their activities on the civilian population and civilian objects"); *id.* r. 115 cmt. ¶¶ 2, 4–5, at 478–79 (verification of targets, based on AP I art. 57(2)(a)(i); rule applies specifically to those who "plan or decide upon" attacks; "feasible" means "practicable or practically possible, taking into account all circumstances ruling at the time"); *id.* r. 116 cmt. ¶ 5, at 480–81 (choice of means, based on AP I art. 57(2)(a)(ii); duty to minimize "indirect as well as direct collateral damage" given "interconnectivity of cyber infrastructure, particularly between military and civilian systems"); *id.* r. 117 cmt. ¶ 2, at 481 (proportionality, AP I art. 57(2)(a)(iii); "continuing personal obligation"); *id.* r. 118 cmt. ¶¶ 5–6, at 482 (choice of targets, AP I art. 57(3); applies where alternative target achieves "comparable military effects"); *id.* r. 119 cmt. ¶ 7, at 484 (cancellation, AP I art. 57(2)(b); "an attacker who initiates a cyber attack has a duty to monitor the attack as long as it is feasible to do so"; difficulty of monitoring "would heighten the degree of scrutiny that is merited during the planning and decision phases of the attack"). Rule 114 cmt. ¶ 1, at 476 records the experts' conclusion as to customary status; individual Rules note dissenting or qualified expert positions where they exist. The Manual states that it "reflects the experts' views" and does not represent the official position of any government or intergovernmental organization. *Id.* Introduction, at 3–4.

institutional precision, what the term feasible means when the targeting system is an algorithm whose decision logic is opaque - specifically, that feasible verification requires pre-deployment testing, training-data auditing, and discrimination-reliability certification before any commander's authorization can be considered legally informed, a reading that aligns directly with Rule 115's confirmation that the obligation extends to those who "plan or decide upon" attacks and not merely to those who execute them. The in-operandi auditability component operationalizes the attack-suspension obligation of Rule 119: it defines the technical architecture - a hard interrupt protocol at the mission-controller layer - through which suspension remains physically possible after deployment authorization but before the weapons effect is consummated; Rule 119's commentary expressly acknowledges that "some cyber-attacks may be difficult to continuously monitor, thus making it practically difficult to know whether to cancel or suspend them" and that this difficulty "would heighten the degree of scrutiny that is merited during the planning and decision phases of the attack," which is precisely the epistemic burden that the interpretability component is designed to discharge. The ex post forensic recoverability component operationalizes the accountability obligations implicit in Rule 114 of Tallinn Manual 2.0: the requirement of "constant care" throughout cyber operations, which presupposes that precautionary measures remain assessable after execution, and creates a chain of responsibility across all persons involved in operations. Rule 118's obligation to select targets causing the least civilian danger among comparable military objectives applies to "all persons who are involved in target selection, approval, and execution," a formulation that presupposes reconstructible decision logic, since accountability for target selection requires that the selection basis remain accessible after the fact. Together, these rules presuppose that targeting decisions remain reconstructible after execution, a condition that design-inherent opacity architecturally forecloses in Tier 3

systems, confirming that the recoverability component is not an additional burden but the minimum epistemic architecture that existing IHL obligations already demand. The ATO's mandatory pre-deployment algorithmic audit and post-engagement logging obligation supply that mechanism.

These three components are not confined to AP I as treaty law. The ICRC Customary IHL Study identifies the obligation to take all feasible precautions in attack as Rule 15, applicable in both international and non–international armed conflict, and the ICJ has confirmed in *Nicaragua* and *Nuclear Weapons* that customary rules emerge from convergent State practice and *opinio juris* under a two–element methodology. To the extent that the ATO is asserted against States not party to AP I, such as the United States in the Project Maven context, its normative force rests on the customary character of the underlying precautionary obligations reflected in ICRC Customary IHL r. 15. While a comprehensive two-element demonstration of the ATO's customary status is beyond the scope of this chapter, this dissertation proceeds on the basis that the core precautionary obligation, to do everything feasible to verify that targets are military objectives, has been recognized as customary international law by the ICJ in *Nuclear Weapons* and codified in the ICRC Customary IHL Study.[207] To the extent that the ATO's three specific components exceed what customary law currently requires, the framework applies as treaty law to AP I States Parties and as an Article 31(1) VCLT interpretive claim as to the content of customary precautionary duties.[208]

207 A further research avenue concerns the application of Common Article 3 and AP II to autonomous cyber weapons in non-international armed conflict, including the horizontal applicability of IHL obligations to non-State armed groups as both targets and potential deployers of autonomous systems; this question is reserved for separate treatment.

208 Jean-Marie Henckaerts & Louise Doswald-Beck, Customary International Humanitarian Law r. 15, at 51–53 (Cambridge Univ. Press 2005); Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 1996 I.C.J. Rep. 226, ¶¶ 78–79 (July 8); *Nicaragua*, *supra* note 35, ¶¶ 183–186.

8. The Algorithmic Transparency Obligation in State Practice: Three Documented Operational Breaches

Three documented State deployments, Turkey's Kargu–2 in Libya,[209] Russia's Lancet–3 in Ukraine,[210] and the United States' Project Maven programme,[211] demonstrate that the ATO's requirements have been systematically breached in contemporary armed conflict, providing empirical grounding for the chapter's structural incompatibility thesis.[212]

a) The STM Kargu–2 in Libya (2020): Severance of In–Operandi Auditability and the Defeat of Article 57(2)(b)

The clearest empirical instantiation of an in–operandi auditability breach arises from the deployment of the Turkish STM Kargu 2 loitering munition system during the Libyan conflict in 2020. The Panel of Experts established pursuant to Security Council resolution 1973 (2011) reported that forces affiliated with the Government of National Accord employed lethal autonomous weapon systems,[213] including the Kargu 2, against logistics convoys and retreating units of the Haftar–aligned forces. The Panel recorded that these systems "were programmed to attack targets without requiring data connectivity between the operator and the munition," characterizing this as a fire, forget, and find

209 Panel of Experts on Libya, *Final Report of the Panel of Experts on Libya Established Pursuant to Security Council Resolution 1973 (2011)*, ¶ 63, U.N. Doc. S/2021/229 (Mar. 8, 2021), https://documents.un.org/doc/undoc/gen/n21/037/72/pdf/n2103772.pdf (last visited Apr. 16, 2026) [hereinafter Libya POE Report].

210 Office of the U.N. High Commissioner for Human Rights, *Report on the Human Rights Situation in Ukraine: 1 September–30 November 2024*, ¶¶ 36–37 (Dec. 31, 2024), https://ukraine.ohchr.org/sites/default/files/2024-12/PR41%20Ukraine%202024-12-31.pdf (last visited Apr. 16, 2026) [hereinafter OHCHR, *Ukraine Report Dec. 2024*].

211 U.S. Dep't of Defense, Memorandum from Deputy Secretary of Defense Robert O. Work, *Establishment of an Algorithmic Warfare Cross-Functional Team (Project Maven)* (Apr. 26, 2017), https://media.defense.gov/2017/Apr/26/2001733373/-1/-1/0/ESTABLISHMENT-OF-AN-ALGORITHMIC-WARFARE-CROSS-FUNCTIONAL-TEAM.PDF (last visited Apr. 16, 2026) [hereinafter DoD, *Project Maven Memo*].

212 Libya POE Report, *supra* note 209, ¶¶ 63–64; U.N. Human Rights Council, *Report of the Independent International Commission of Inquiry on Ukraine*, ¶ 23, U.N. Doc. A/HRC/52/62 (Mar. 15, 2023), https://documents.un.org/doc/undoc/gen/g23/049/73/pdf/g2304973.pdf (last visited Apr. 16, 2026) [hereinafter HRC, *Ukraine Inquiry Report*]; DoD, *Project Maven Memo*, *supra* note 211, at 1.

213 Libya POE Report, *supra* note 209, ¶ 63 ("Logistics convoys and retreating HAF were subsequently hunted down and remotely engaged by the unmanned combat aerial vehicles or the lethal autonomous weapons systems such as the STM Kargu-2 . . . and other loitering munitions. The lethal autonomous weapons systems were programmed to attack targets without requiring data connectivity between the operator and the munition: in effect, a true 'fire, forget and find' capability.").

capability and identifying it as the first reported instance of an autonomous weapon system conducting lethal engagements in an active armed conflict without maintained operator oversight.[214]

The Libyan deployment also engages the ex–post forensic recoverability component. The Panel report does not specify the individual targeting decisions made by Kargu-2 systems, a gap that is consistent with, and this dissertation contends results from, the architectural absence of decision-logging mechanisms capable of preserving the system's classification rationale at the moment of each engagement. This forensic opacity does not merely complicate post hoc accountability; it impedes the attribution analysis through which State responsibility under Article 4, 7, 8, and 14(3) of the ARSIWA on State responsibility, and potential individual criminal responsibility under the Rome Statute, would otherwise be examined.

b) The ZALA Lancet–3 in Ukraine (2022–present): Adversarial Distribution Shift and the Defeat of Ex-ante Interpretability under AP I

The Russian Lancet 3 loitering munition, developed by ZALA Aero Group and deployed by Russian armed forces in Ukraine from late 2022 onwards, provides a prominent empirical manifestation of the distribution–shift failure mode identified in Section II.6 above within an ongoing international armed conflict between High Contracting Parties to AP I. Open–source documentation indicates that the Lancet 3 employs electro–optical and television guidance systems integrated with AI–driven pattern–recognition algorithms, powered by NVIDIA's Jetson TX2 module, enabling autonomous object tracking and semi–autonomous or autonomous terminal–phase target acquisition without continuous operator input. Visually confirmed strike records show engagement of main battle tanks, self–propelled artillery,

214 Libya POE Report, *supra* note 209, ¶ 63.

surface–to–air missile systems, radar vehicles, and infrastructure installations, a target set encompassing objects that are military objectives by nature and objects whose status depends on context–specific application of Article 52(2)'s "circumstances ruling at the time" standard.[215]

The Independent International Commission of Inquiry on Ukraine has documented a consistent pattern of failure by Russian armed forces to take feasible precautions to verify the presence of civilians and civilian objects and has found indiscriminate and disproportionate attacks in violation of international humanitarian law. From an ATO perspective, this evidentiary record is analytically significant: it constitutes institutional confirmation at the level of a United Nations Human Rights Council inquiry mechanism that ex post reconstruction of the decision–making underlying numerous Russian strike patterns is not available from the public record, underscoring that the forensic recoverability component of the ATO is not satisfied in such operations. At the same time, Ukrainian counter–measures, including camouflage of armored vehicles, rapid relocation of artillery and the adjustment of electromagnetic signatures of mobile air defense assets, have systematically altered the operational feature distribution on which pre-deployment testing and training of Russian loitering munitions and other remotely operated systems were based, a dynamic consistent with, and on the analysis advanced in this dissertation constituting, the adversarial distribution shift that renders pre-deployment confidence metrics unreliable as a proxy for the "everything feasible" standard under Article 57(2)(a)(i) AP I verification requirement. Where a system's classification architecture has been trained on a feature distribution that the adversary has deliberately rendered obsolete,

215 *Lancet Loitering Munition System, Russia*, Army Technology (Sept. 11, 2024), https://www.army-technology.com/projects/lancet-loitering-munition-system-russia/ (last visited Apr. 16, 2026); David Hambling, *Russian Loitering Munition Racks Up Kills But Shows Limitations*, Forbes (Dec. 1, 2022), https://www.forbes.com/sites/davidhambling/2022/12/01/russian-loitering-munition-racks-up-kills-but-shows-limitations/ (last visited Apr. 16, 2026) [hereinafter Hambling, *Lancet Limitations*].

commanders cannot credibly determine that everything feasible has been done to confirm that the system will reliably distinguish military objectives from civilian objects "in the circumstances ruling at the time" of attack.[216]

c) United States Project Maven: Specification Gaming and the Structural Incapacity of the Article 52(2) Military Objective Determination

The United States' Project Maven, the Algorithmic Warfare Cross-Functional Team established in 2017 to operationalize artificial intelligence for object recognition in full-motion video within the Department of Defense, constitutes the most intensively documented example of the structural tension between proxy-metric optimization and the legal military-objective determination required by Article 52(2) AP I in a State AI-enabled targeting programme, demonstrating that the specification-gaming risk is present even in systems retaining human review if the legal military-objective determination is in practice delegated to, and bounded by, the algorithmic confidence score.[217] Publicly released defense materials and independent analyses confirm that Project Maven employs deep–learning convolutional neural networks to classify objects and generate confidence–scored labels used to inform human targeting decisions, with fielding decisions accepted at significantly below perfect classification accuracy on the principle of deploying the most capable available system rapidly.[218]

Project Maven's optimization objective, improving classification accuracy against labelled training data, thus exemplifies specification gaming: the system optimizes performance against a proxy objective

216 HRC, *Ukraine Inquiry Report*, *supra* note 212, ¶¶ 23–27, 32–34, 43–44; AP I, arts. 51(4)–(5), 57(2)(a)(i).

217 DoD, *Project Maven Memo*, *supra* note 211, at 1.

218 DoD, *Project Maven Memo*, *supra* note 211, at 1; Alice Spazian, Arthur Holland Michel & Alisha Anand, *UNIDIR on Lethal Autonomous Weapons: Mapping Our Research to the Discussions of the GGE on LAWS* 5–7 (UNIDIR 2023), https://unidir.org/wp-content/uploads/2023/05/UNIDIR-on-Lethal-Autonomous-Weapons-Final.pdf (last visited Apr. 16, 2026) [hereinafter Spazian, Michel & Anand, *UNIDIR on LAWS*].

(pattern similarity) that systematically diverges from the legal objective (military–objective determination under Article 52(2) AP I). The ATO's *ex ante* interpretability component requires that commanders understand this divergence and retain the ability to verify, independently of the system's confidence scores, that targets selected for engagement satisfy the cumulative criteria of Article 52(2) AP I in the actual circumstances of attack; opaque deep–learning architectures that do not furnish meaningful insight into their decision logic frustrate this requirement.[219]

d) Doctrinal Synthesis: Cumulative Evidentiary Grounding of the ATO and the Threshold for an Article 36 Prohibitory Conclusion

The three deployments examined, Kargu–2, Lancet–3 and Project Maven, instantiate each of the ATO's components in documented operational contexts. Kargu–2 demonstrates a structural breach of in–operandi auditability through intentional severance of connectivity and consequent architectural impossibility of discharging the cancellation–and–suspension duty; Lancet–3 demonstrates the defeat of *ex-ante* interpretability under adversarial distribution shift in an active international armed conflict; and Project Maven demonstrates substitution of proxy confidence metrics for the legal determination required by Article 52(2). Together, these cases show that the ATO is not a speculative construct extrapolated from technical literature but the legal distillate of recurrent patterns of breach of existing precautionary and weapons–review obligations across distinct State deployments.[220]

States that deploy autonomous targeting systems without satisfying the ATO's three components incur a breach of the precautionary obligation at the moment of deployment, engaging State responsibility under the ARSIWA framework and, where ICC jurisdiction is present, potential

219 AP I, art. 52(2); Spazian, Michel & Anand, *UNIDIR on LAWS*, *supra* note 218, at 7.

220 Libya POE Report, *supra* note 209, ¶ 63; HRC, *Ukraine Inquiry Report*, *supra* note 212, ¶¶ 23, 32–33.

individual criminal responsibility — the accountability framework developed in Chapter 4.

III. Special Protections: Medical Units, Objects Indispensable to Survival, and Works Containing Dangerous Forces

The structural verification deficits established in Section II — temporal disaggregation, cascade unpredictability, and machine-learning opacity — are categorically prohibitive rather than merely heightened when autonomous cyber weapons operate against objects to which AP I assigns special protection.

AP I establishes heightened protection for certain categories of civilian objects whose disruption poses especially grave humanitarian risks. These special protections for medical units, Article 12 AP I, objects indispensable to survival, Article 54 AP I, and works containing dangerous forces, Article 56 AP I, impose more stringent verification duties that, for autonomous cyber weapons, create acute compliance challenges because the functional assessments required demand semantic understanding and real-time contextual awareness that autonomous systems deployed as of the date of this dissertation lack.[221]

1. Medical Units and Digitized Healthcare Systems under Article 12 AP I

Modern hospitals rely on networked medical devices, electronic health records systems, and building automation systems that may use identical hardware and protocols to non–medical commercial systems. An autonomous targeting algorithm scanning for "vulnerable Windows servers" or "industrial control systems" cannot distinguish, based on network traffic patterns alone, between a hospital's HVAC control system and a commercial office building's climate control. Autonomous propagation through shared vulnerabilities cannot discriminate by

221 AP I, arts. 12, 54, 56; 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 520, 2154.

protected status, violating Article 12 when foreseeable effects reach medical units regardless of targeting intent.[222] State practice during the COVID–19 pandemic and the 2022–2025 Ukraine conflict demonstrates that cyber operations targeting healthcare infrastructure constitute attacks under Article 49(1) AP I when they render medical facilities incapable of performing their humanitarian functions.[223]

The ICRC characterized such operations as violations of IHL when conducted during armed conflict, stating that "belligerents must respect and protect medical facilities and personnel at all times," which includes ensuring that "there should be no interference with their work" and that "the obligation to respect and protect medical facilities encompasses a prohibition against deleting, altering, or otherwise negatively affecting medical data.[224] Ransomware attacks that encrypt medical records, disable medical devices, or disrupt hospital IT systems will, where they are designed to render a hospital's computer systems or networks non–functional, fall within the notion of "attacks" under Article 49(1) AP I on the ICRC's functional–harm interpretation of cyber operations. When such operations render a hospital operationally incapable of treating patients, their humanitarian consequences are comparable to those of physically destroying the facility and engaging the protections of Article 12 AP I and Article 19(1) GC I, under which medical units must 'at all times' be respected and protected and "shall not be the object of attack." For autonomous or highly automated cyber capabilities, these

222 AP I, arts. 12, 48, 54(1); 1987 ICRC AP Commentary, *supra* note 18, ¶ 517; Tallinn Manual 2.0, *supra* note 18, r. 92 cmt. ¶¶ 6–7, r. 93 cmt. ¶ 3; Schmitt, *Reply to the Critics*, *supra* note 151, at 11.

223 AP I, art. 49(1); Tallinn Manual 2.0, *supra* note 18, r. 92 cmt. ¶ 2, r. 131 cmt. ¶ 5; Kubo Mačák, Tilman Rodenhäuser & Laurent Gisel, *Cyber Attacks Against Hospitals and the COVID-19 Pandemic: How Strong Are International Law Protections?*, ICRC Humanitarian L. & Pol'y Blog (Apr. 2, 2020), https://blogs.icrc.org/law-and-policy/2020/04/02/cyber-attacks-hospitals-covid-19-pandemic-how-strong-international-law-protections (last visited Apr. 16, 2026) [hereinafter Mačák, Rodenhäuser & Gisel, *Cyber Attacks Against Hospitals*].

224 Laurent Gisel, Tilman Rodenhäuser & Knut Dörmann, *Twenty Years On: International Humanitarian Law and the Protection of Civilians Against the Effects of Cyber Operations During Armed Conflicts*, 102 Int'l Rev. Red Cross 287, 327–28 (2020) [hereinafter Gisel, Rodenhäuser & Dörmann, *Twenty Years On*] ("Given the fundamental importance of health care for anyone affected by armed conflict, belligerents must respect and protect medical facilities and personnel at all times . . . the obligation to respect and protect medical facilities encompasses a prohibition against deleting, altering or otherwise negatively affecting medical data."); *accord* Mačák, Rodenhäuser & Gisel, *Cyber Attacks Against Hospitals*, *supra* note 223.

risks are heightened given the acknowledged difficulty of ensuring reliable distinction in complex, interconnected digital infrastructure, including health–care systems, when systems select and attack targets based on pre–programmed profiles and sensor inputs. The ICRC's updated Commentary on Article 19 GC I confirms that medical establishments and units enjoy continuous, non–reciprocal protection 'at all times', and, taken together with the obligation in Article 57(2)(a)(i) AP I to do everything feasible to verify that a target is not a protected person or object, requires attackers to verify that a facility is not a medical unit, or that its protection has lawfully ceased, before conducting cyber operations expected to render it dysfunctional.[225] Where algorithmic opacity precludes commanders from understanding how autonomous targeting systems classify network assets, and where temporal disaggregation separates initial network intrusion from effects activation by months or years, the requirement under Article 57(2)(a)(i) AP I to verify that objectives are military targets becomes structurally impossible to satisfy in real time. The integration of civilian healthcare infrastructure with commercial cloud services and shared IT platforms further complicates verification: disruption of Amazon Web Services or Microsoft Azure infrastructure used by military logistics systems may simultaneously render inoperable civilian hospital electronic health records, medical device connectivity, or telemedicine platforms hosted on the same commercial infrastructure, violating Article 12 through cascade effects even when medical facilities were not directly targeted.

225 Geneva Convention for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field art. 19(1), Aug. 12, 1949, 75 U.N.T.S. 31 [hereinafter GC I] ("shall at all times be respected and protected"); AP I, arts. 12(1), 49(1); 1987 ICRC AP Commentary, *supra* note 18, ¶ 520; Int'l Comm. of the Red Cross, *Commentary on the First Geneva Convention*, art. 19, ¶ 1809 (2d ed. 2016) [hereinafter ICRC GC I Commentary], https://ihl-databases.icrc.org/en/ihl-treaties/gci-1949/article-19/commentary (last visited Apr. 16, 2026); Gisel, Rodenhäuser & Dörmann, *Twenty Years On*, *supra* note 224, at 287, 294, 328.

2. Objects Indispensable to Civilian Survival under Article 54 AP I

Autonomous cyber weapons targeting "industrial control systems" or "SCADA networks" face a profound identification challenge: a Siemens PLC controlling a centrifuge at a military research facility may be functionally and technically identical to a PLC controlling pumps at a municipal water treatment plant. Without a contextual understanding of the facility's purpose, information not reliably extractable from network traffic patterns or device configurations, an autonomous weapon cannot satisfy Article 57(2)(a)(i)'s AP I verification requirement before attacking.[226]

The 2015 and 2016 cyber-attacks on Ukraine's electricity distribution infrastructure demonstrated how cyber operations can render essential civilian infrastructure dysfunctional at scale, engaging Article 54's AP I prohibition on attacks against objects indispensable to survival. In December 2015, cyber attackers gained remote access to supervisory control and data acquisition (SCADA) systems at three Ukrainian electricity distribution companies, opening circuit breakers and cutting power to approximately 230,000 civilians for several hours in sub–zero winter temperatures.[227] The December 2015 attacks targeted distribution–level infrastructure at three Ukrainian power companies, demonstrating sophistication in manipulating industrial control systems to remotely open circuit breakers at 27 substations. While these operations targeted electricity infrastructure rather than water or food systems directly, they illustrate a potential functional harm concern

226 AP I, arts. 54(2), 57(2)(a)(i); 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 2100–2101; Tallinn Manual 2.0, *supra* note 18, r. 115 cmt. ¶ 5; Schmitt, *Reply to the Critics*, *supra* note 151, at 11.

227 AP I, art. 54(2); 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 2089–2105; ICS-CERT, *Alert (ICS-ALERT-16-056-01): Cyber-Attack Against Ukrainian Critical Infrastructure* (Feb. 25, 2016), https://www.cisa.gov/news-events/ics-alerts/ics-alert-16-056-01 (last visited Apr. 16, 2026) [hereinafter ICS-CERT, *Ukraine Alert*]; Electricity Information Sharing & Analysis Ctr. & SANS Indus. Control Sys., *Analysis of the Cyber Attack on the Ukrainian Power Grid: Defense Use Case*, at iv, 8 (Mar. 18, 2016), https://www.nerc.com/pa/CI/ESISAC/Documents/E-ISAC_SANS_Ukraine_DUC_18Mar2016.pdf (last visited Apr. 16, 2026) [hereinafter E-ISAC & SANS, *Ukraine Power Grid*.

under IHL: cyber operations that render essential civilian services inoperable may implicate the prohibition under Article 54(2) AP I on "rendering useless" objects indispensable to survival, particularly where such operations indirectly deprive the population of water or food supplies during armed conflict.[228]

Article 54(3) AP I permits derogation from the prohibition only for objects used "in direct support of military action," such as bombarding a food–producing area to prevent enemy advance or attacking a storage barn used as an arms depot. This exception cannot justify autonomous cyber weapons targeting essential civilian infrastructure because such systems lack the contextual judgment required to assess whether the object's specific use constitutes "direct support" as narrowly defined by the ICRC Commentary. Moreover, Article 54(3) AP I prohibits any destruction of objects covered by Article 54(2) whose purpose or effect is to starve civilians as a method of warfare or to cause them to move away for that reason, a prohibition that admits no derogation even under military necessity where the protected objects are not in direct support of adverse military action.[229] Autonomous weapons operating through feature–based classification cannot make these context–dependent, necessity–calibrated determinations. A cyber operation that disables water treatment infrastructure serving both military bases and civilian populations cannot satisfy Article 54's AP I heightened verification standard through algorithmic confidence scores; it requires human judgment that integrates knowledge of civilian dependency, alternative targeting options, and proportionality considerations that autonomous systems cannot replicate. Where cyber operations targeting dual–use infrastructure create reasonably foreseeable risks of civilian

228 ICS-CERT, *Ukraine Alert*, *supra* note 1, at 1; E-ISAC & SANS, *Ukraine Power Grid*, *supra* note 1, at iv–v; Francesca Capone, *Cyber Operations Against Critical Infrastructure and the Prohibition on Attacking Objects Indispensable to the Survival of the Civilian Population*, 102 Int'l Rev. Red Cross 731, 750 (2020) [hereinafter Capone, *Cyber Operations*]; 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 2100–2103, 2110–2111; Tallinn Manual 2.0, *supra* note 18, r. 81 cmt. ¶¶ 2–3; AP I, arts. 54(2)–(3).

229 AP I, art. 54(3).

starvation, disease outbreak, or deprivation of essential services, the precautionary principle embodied in Article 57 AP I requires that uncertainty be resolved in favor of civilian protection, prohibiting the deployment of autonomous weapons whose targeting decisions cannot be verified as satisfying Article 54's AP I stringent standards.[230]

Article 56's protection of works containing dangerous forces, dams, dykes, and nuclear electrical generating stations applies in principle to cyber operations targeting such installations, though the detailed analysis of nuclear-facility-specific scenarios falls outside the scope of this study's central thesis. Article 56 AP I prohibits attacks against works containing dangerous forces even where they constitute military objectives if the attack may cause severe losses among the civilian population. For autonomous cyber weapons, the structural verification deficit is most acute here: a system incapable of distinguishing a dam's civilian water-management function from a military logistics server cannot satisfy the heightened verification burden that Article 56's absolute protection imposes. The cascade "*verification impossibility"* analysis of Section II.3 applies with compounded force to Article 56 objects, where the consequences of a classification error are potentially catastrophic and irreversible.

The final challenge concerns commercial cyber infrastructure, cloud computing platforms, commercial data centers, and internet backbone routers that serve both military and civilian users simultaneously. Under the presumption under Article 52(3) AP I, such objects must be treated as civilian unless their military use can be definitively confirmed.[231] Moreover, because this infrastructure supports hospitals, water treatment facilities, emergency services, and financial systems, attacks on commercial infrastructure predictably engage the special protection

230 AP I, arts. 57(2)(a)(i), 58; 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 2194–2195, 2244.

231 AP I, arts. 52(2)–(3), 56(1), 57(2)(a)(i); 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 2022, 2028, 2032, 2153; Tallinn Manual 2.0, *supra* note 18, r. 115 cmt. ¶ 5; Schmitt, *Reply to the Critics*, *supra* note 151, at 16–17.

categories indirectly through cascade effects. A cyber operation targeting cloud infrastructure used by military logistics systems may simultaneously disrupt civilian hospital databases hosted on the same platform, water utility management systems that depend on the same network connectivity, or emergency services dispatch systems that use the same commercial services. The Article 57(2)(a)(ii) AP I duty to "take all feasible precautions in the choice of means and methods of attack with a view to avoiding, and in any event to minimizing, incidental loss of civilian life" requires attackers to assess these reverberating effects and select means that minimize harm to specially protected objects.[232]

Temporal disaggregation, algorithmic opacity, and cascade verification impossibility, together, not independently, produce *verification impossibility* as a legal consequence; no single feature, standing alone, generates the categorical prohibition.

The per se prohibition attaches to systems exhibiting all three features concurrently; a system lacking any one of the three structural characteristics identified there falls outside the categorical prohibition, though it may still attract heightened precautionary obligations under Article 57 AP I. The appropriate response is not enhanced interpretive guidance for existing rules, but a preventive governance architecture whose design is the subject of Chapters 5 and 6.

Chapter 3 has established that AP I's precautionary rules cannot govern Tier 3 autonomous cyber weapons. The question that follows is: who is legally responsible when those rules are violated? That is the question Chapter 4 answers.

232 AP I, arts. 12(1), 54(2), 57(2)(a)(ii); 1987 ICRC AP Commentary, *supra* note 18, ¶ 2200; Tallinn Manual 2.0, *supra* note 18, r. 116 cmt. ¶ 5.

Chapter 4
Accountability and the Structural Responsibility Gap

When IHL's precautionary architecture cannot operate — because the human it presupposes has been eliminated by design — a second question immediately follows: who is legally responsible?

The answer that accountability law offers is, on its face, straightforward: the commander who authorized the weapon, the State whose forces deployed it, the individual who ordered its use.

This Chapter demonstrates that none of those answers works as the law currently stands — and that understanding precisely why they fail is the only path to a solution that does.

The structural defect identified in Chapters 2 and 3 — no human present at the moment of lethal effect — defeats accountability in the same way it defeats precaution. ARSIWA's effective control standard cannot attach where targeting decisions are algorithmic and the causal chain has no human at its operative moment. Article 28 of the Rome Statute's command responsibility doctrine cannot generate the "should have known" knowledge standard where the system's architecture constitutively forecloses what any human supervisor could know.

The result is a structural responsibility gap: not a gap in the law, remediable by analogy, but a categorical mismatch between accountability law's foundational presuppositions and the operational architecture of Tier 3 autonomous cyber weapons.

This Chapter's response is a temporal reorientation. The accountability inquiry cannot be made to work at the operational phase — the moment of autonomous targeting — because no human is present. But it can be made to work at the deployment authorization stage, where a human

being still exists, still decides, and still bears legal responsibility for what that decision sets in motion. A commander who knew or "should have known", at the moment of authorization, that the system's architecture would produce IHL-violating conduct, and who failed to conduct the Article 36 AP I weapons review that was the necessary and reasonable measure to prevent it, satisfies both elements of Article 28(1)(a) of the Rome Statute.

This is not new law. It is existing law applied at the correct temporal locus.

Section I establishes the normative framework and develops the deployment-stage reorientation through the command responsibility jurisprudence of the ICTY and ICC.

Section II identifies the structural remedies that the accountability gap demands — and explains why they are legal necessities, not policy options.

I. Command Responsibility and Autonomous Cyber Weapons: Structural Obstacles and the Authorization-Stage Reorientation

The doctrine of command responsibility was developed to close the accountability gap that the perpetrator model leaves open: where harmful acts are committed not by the commander personally but by subordinates operating within a chain of command that the superior controls, the law imposes criminal liability for failure to prevent or repress those acts precisely because the superior's authority creates both the capacity and the obligation to intervene. That foundational logic — superior authority generating superior accountability — encounters a structural obstacle in the context of autonomous cyber weapons that is qualitatively distinct from any previously encountered in the command responsibility jurisprudence of the ICTY and ICC: the subordinate whose conduct triggers liability is not a human agent whose acts can be monitored, anticipated, and interrupted, but an algorithmic system whose targeting

decisions are constitutively opaque to any human supervisor possessing only the means that current technology makes available. The sections that follow demonstrate that this opacity defeats the command responsibility framework at two levels — operationally, because the "should have known" standard of Article 28 Rome Statute and Article 86(2) of AP I cannot be satisfied in respect of conduct that is unknowable using any means available to human commanders; and structurally, because the accountability gap that operational opacity creates cannot be closed by reformulating the standard of knowledge, but only by relocating the accountability inquiry from the operational phase to the authorization stage, where human decision–making remains genuinely operative and the commander's obligation to create adequate informational conditions for command is capable of being discharged. Section I.1 below establishes the normative framework of Article 86 AP I and Article 28 of the Rome Statute, as interpreted through the ICTY and ICC jurisprudence, and identifies the precise points at which autonomous cyber weapons defeat the agency presuppositions those instruments share. Section I.2 below develops the opacity obstacle in technical and doctrinal detail, demonstrating that the "black box" character of deep neural network architectures is not a contingent engineering limitation but a constitutive feature of the deep learning paradigm that no supervisory protocol can overcome at the operational phase. Section I.3 below advances this chapter's original doctrinal contribution: the reorientation of the command responsibility inquiry to the authorization stage, grounding liability in the commander's decision to deploy a system structurally incapable of generating the oversight information that the obligation to create adequate informational conditions for command, as reflected in Article

28(1)(a) Rome Statute and the ICTY's constructive knowledge jurisprudence, presupposes.[233]

1. The Normative Framework: Article 86 AP I and Article 28 Rome Statute

The doctrine of command responsibility provides a basis for individual criminal liability that operates independently of the direct perpetrator model and is, in principle, better suited to contexts in which harmful acts are committed by persons not directly controlled by the accused. Article 86(2) of AP I imposes liability on superiors who fail to take 'all feasible measures within their power' to prevent or repress breaches by subordinates where the superior 'knew, or had information which should have enabled him to conclude in the circumstances at the time, that' the subordinate was about to commit or had committed a breach.[234] Article 28 Rome Statute adopts parallel requirements: a military commander is criminally responsible for crimes committed by forces under effective command and control where the commander "knew or, owing to the circumstances at the time, should have known" that the forces were committing or about to commit such crimes (Article 28(1)(a)), and failed to take all necessary and reasonable measures to prevent or repress their commission (Article 28(1)(b)).[235]

The ICTY Appeals Chamber in *Hadžihasanović* confirmed that command responsibility in non-international armed conflict attaches to any commander of an organized military force who knew or had reason to know that subordinates were committing or were about to commit

233 Rome Statute of the International Criminal Court, art. 28(1)(a), *opened for signature* July 17, 1998, 2187 U.N.T.S. 90, https://www.icc-cpi.int/sites/default/files/RS-Eng.pdf (last visited Apr. 16, 2026) [hereinafter Rome Statute] ("owing to the circumstances at the time, "should have known"" that forces were committing or about to commit crimes); Prosecutor v. Jean-Pierre Bemba Gombo, Case No. ICC-01/05-01/08-3343, Trial Chamber III, Judgment pursuant to Article 74 of the Statute, ¶ 170(d) (Int'l Crim. Ct. Mar. 21, 2016) [hereinafter Bemba Trial Judgment] (restating elements of art. 28(1)); *id.* ¶ 196; Prosecutor v. Jean-Pierre Bemba Gombo, Case No. ICC-01/05-01/08, Appeals Chamber, Judgment, ¶ 173 (Int'l Crim. Ct. June 8, 2018) [hereinafter Bemba Appeals Judgment] (reversing on adequacy of measures assessment, not the epistemic threshold of art. 28(1)(a)). This dissertation relies on Bemba Trial Judgment ¶ 170(d) and Rome Statute, *supra* note 233, art. 28(1)(a) as the operative *lex lata*.

234 AP I, arts. 86(1)–(2); Rome Statute, *supra* note 233, art. 28(1).

235 Rome Statute, *supra* note 233, arts. 28(1)(a)–(b).

violations and failed to take the necessary and reasonable measures to prevent or punish them, thereby confirming the affirmative and constructive character of the effective command obligation.[236] Under Article 86(2) AP I, the constructive knowledge standard attaches where a commander "had information which should have enabled him to conclude in the circumstances at the time" that a subordinate was committing a breach, a formulation predicated on the possession and activation of informational means. The ICC Trial Chamber III established, in respect of the correlative duty to prevent under Article 28(1)(b), that its scope depends on the material power of the commander to intervene, encompassing institutional obligations, including ensuring adequate IHL training, securing compliance reports, issuing remedial orders, and taking disciplinary measures, through which the conditions for effective oversight of subordinate conduct are maintained.[237] This formulation will prove important to the doctrinal reorientation advanced in Section I.3 below.

236 Prosecutor v. Hadžihasanović & Kubura, Case No. IT-01-47-A, Appeals Chamber, Judgment, ¶¶ 30, 33 (Int'l Crim. Trib. for the Former Yugoslavia Apr. 22, 2008) [hereinafter Hadžihasanović AC Judgment]; Prosecutor v. Hadžihasanović, Alagić & Kubura, Case No. IT-01-47-AR72, Decision on Interlocutory Appeal Challenging Jurisdiction in Relation to Command Responsibility, ¶ 18 (Int'l Crim. Trib. for the Former Yugoslavia July 16, 2003) [hereinafter Hadžihasanović Interlocutory Appeal] (confirming that command responsibility is applicable in non-international armed conflict as a matter of customary international law and that the obligation is affirmative in character, derived from the requirements of responsible command); Prosecutor v. Delalić et al., Case No. IT-96-21-T, Judgment, ¶¶ 333–334 (Int'l Crim. Trib. for the Former Yugoslavia Nov. 16, 1998) [hereinafter Čelebići Judgment] (establishing that international law imposes an "affirmative duty" on superiors to prevent violations by subordinates as the doctrinal basis of command responsibility under art. 7(3)).

237 AP I, art. 86(2); Bemba Trial Judgment, *supra* note 233, ¶ 203 (holding that the scope of the duty to take all necessary and reasonable measures within the commander's power is intrinsically linked to the commander's material ability to prevent or repress crimes, and that such measures encompass institutional obligations-including adequate IHL training, the securing of compliance reports, the issuance of remedial orders, and the imposition of disciplinary measures-through which the conditions for effective oversight of subordinate conduct are maintained).

2. The Black Box Problem: Opacity as a Structural Legal Obstacle

The command responsibility framework appears, at first examination, to offer considerably more promise than the perpetrator model: the "should have known" standard, as interpreted in *Bemba Gombo*, creates a constructive knowledge obligation that does not require proof of specific intent or actual awareness of particular criminal acts. A commander who deploys an autonomous cyber weapon without establishing adequate monitoring, assessment, and deactivation mechanisms fails precisely the *Bemba Gombo* obligation to create informational conditions for effective command: the failure is not passive ignorance but an active, structural choice to deploy a system whose architecture guarantees ignorance of its specific targeting decisions.

The structural obstacle is, however, equally fundamental. Commanders cannot exercise "should have known" oversight of conduct that is constitutively unknowable using any means available to human supervisors. Deep neural networks, the architectural basis of most advanced autonomous targeting systems, generate their outputs through learned statistical associations in high–dimensional parameter spaces that do not, absent dedicated transparency and interpretability measures, aggregate into a human–readable decision rationale, even for expert engineers.[238] As the technical literature on machine learning interpretability consistently demonstrates, this opacity is not a temporary engineering limitation awaiting a technical solution: it is a constitutive feature of the deep

238 Nat'l Inst. of Standards & Tech., *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, at 15–16 (Jan. 2023), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf (last visited Apr. 16, 2026) [hereinafter NIST, *AI RMF 1.0*]; Secretary-General, *Lethal Autonomous Weapons Systems: Report of the Secretary-General*, ¶¶ 13–14, 21, U.N. Doc. A/79/88 (June 30, 2024), https://documents.un.org/doc/undoc/gen/n24/154/32/pdf/n2415432.pdf (last visited Apr. 16, 2026) [hereinafter S.G. Report, *LAWS 2024*].

learning paradigm, which achieves its effectiveness at complex pattern–recognition tasks precisely by exceeding human interpretive capacity.[239]

That doctrinal conclusion receives authoritative independent confirmation from the United Nations Institute for Disarmament Research's 2025 evidence-based assessment of military artificial intelligence. UNIDIR identifies opacity as a constitutive technological challenge, concluding that "Modern machine learning (e.g., deep neural networks) often operates in ways that are not explainable to human operators. This lack of transparency makes it hard to assess the trustworthiness of AI systems and to diagnose errors. In military use, an inability to understand why an AI system made a recommendation or took an action can erode human confidence and complicate accountability."[240]

UNIDIR further identifies traceability as a legal obligation rather than a discretionary safeguard, concluding that "traceability in systems or, at the very least, robust documentation and forensic evidence protocols would ensure that States are able to comply with the IHL obligation to conduct effective investigations into alleged violations," an obligation that the constitutive opacity of deep learning architectures structurally forecloses at the operational phase.[241] The UNIDIR assessment thereby confirms, from an independent institutional perspective, the structural

239 Council Regulation 2024/1689 of the European Parliament and of the Council of 13 June 2024 on Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), pmbl. ¶¶ 2, 6, 12, 24, 26, 2024 O.J. (L 2024/1689) [hereinafter EU AI Act], https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689 (last visited Apr. 16, 2026); Nat'l Inst. of Standards & Tech., *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, NIST AI 600-1, at 4–6 (July 2024), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf (last visited Apr. 16, 2026) [hereinafter NIST, Generative AI Profile]; Sassòli, IHL, *supra* note 251, §§ 10.073, 10.088, at 568 (§ 10.073: the "black box" phenomenon renders systems whose outputs cannot be predicted per se prohibited as indiscriminate under AP I art. 51(4)(b), independently of any specific targeting decision; § 10.088: the obligation under AP I art. 57(2)(b) to cancel or suspend an attack that becomes unlawful requires autonomous weapons systems to be constructed to perceive environmental changes "at least as well as humans"—a technical specification that, combined with the constitutive opacity identified in § 10.073, renders Tier 3 systems incapable of satisfying either the oversight or the suspension obligation).

240 UNIDIR Security & Technology Programme, *Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security: An Evidence-Based Road Map for Future Policy Action* 18 (UNIDIR 2025).

241 *Id.* at 18 ("In addition, traceability in systems or, at the very least, robust documentation and forensic evidence protocols would ensure that States are able to comply with the IHL obligation to conduct effective investigations into alleged violations."); *see also* Yana Afina, *Regional Perspectives on the Application of International Humanitarian Law to Lethal Autonomous Weapon Systems* (UNIDIR 2025), https://unidir.org (last visited May 11, 2026).

character of the accountability deficit identified in this Section: the opacity that defeats the "should have known" standard of Article 28(1)(a) Rome Statute at the operational phase simultaneously defeats the IHL investigation obligation at the post-violation phase, rendering both accountability mechanisms structurally unavailable and reinforcing the doctrinal necessity of relocating the responsibility inquiry to the deployment authorization stage.

A commander cannot satisfy the constructive knowledge standard in respect of a system whose decision logic is inaccessible not merely in practice but in principle, because the "information which should have enabled the superior to conclude" — the operative criterion under Article 86(2) AP I and the cognate "had reason to know" standard applied by the ICTY presuppose the existence of informational means capable of generating the relevant knowledge, which the constitutive opacity of deep learning architectures renders structurally unavailable.[242] The responsibility gap thereby reappears within the command responsibility framework, not through absence of authority but through structural impossibility of informed oversight.

242 AP I, art. 86(2) ("if they knew, or had information which should have enabled them to conclude in the circumstances at the time, that he was committing or was going to commit such a breach"); Prosecutor v. Delalić et al. (*Čelebići*), Case No. IT-96-21-A, ICTY Appeals Chamber, Judgment, ¶ 241 (Feb. 20, 2001) ("a superior will be criminally responsible through the principles of superior responsibility only if information was available to him which would have put him on notice of offences committed by subordinates") [hereinafter Čelebići Appeals Judgment]; Prosecutor v. Blaškić, Case No. IT-95-14-A, ICTY Appeals Chamber, Judgment, ¶¶ 62–64 (July 29, 2004) [hereinafter Blaškić Appeals Judgment] (reaffirming Čelebići standard; criminal negligence expressly rejected as basis of liability.

3. Command Responsibility Reoriented: The Authorization Stage

This chapter advances the following original doctrinal proposition as its principal contribution to the command responsibility analysis: the accountability inquiry must be relocated from the operational phase, where autonomous systems preclude contemporaneous human oversight, to the deployment authorization stage, where human decision-making remains genuinely operative. Rather than asking whether the commander "should have known" about specific targeting decisions during operations, a question to which the architecture of Tier 3 systems ensures the answer is structurally negative, the accountability inquiry must center on whether the commander ought to have known, at the time of authorization, that the system's architecture was incapable of generating the oversight information that lawful command requires. This reorientation is consistent with, rather than a departure from, the *Bemba* Gombo framework: the obligation to establish informational conditions for knowledge is logically anterior to the obligation to act on knowledge obtained. The deployment-stage reorientation proposed in this study applies the "should have known" threshold prospectively to the authorization decision, on the basis that a commander who deploys a Tier 3 autonomous cyber weapon without adequate pre-deployment verification of its discrimination capacity "should have known", at the moment of authorization, that the system's architectural incapacity rendered lawful targeting impossible.

The Appeals Chamber's reversal addressed exclusively the Article 28(1)(b) measures element, whether the steps *Bemba* took were adequate given his actual capacity as a remote commander. The reversal did not disturb the Article 28(1)(a) constructive knowledge standard. That standard remains operative as confirmed by the Pre-Trial Chamber at its confirmation decision and constitutes the doctrinal foundation for the deployment-stage reorientation advanced here. The Appeals Chamber at paragraphs 167–170 established that the measures element is

calibrated to the commander's actual material capacity, and at paragraph 189 enumerated seven specific errors all confined to that element, explicitly leaving the Trial Chamber's formulation of the constructive knowledge element undisturbed. This distinction is analytically decisive for this dissertation's deployment-stage reorientation: the reorientation operates exclusively upon the constructive knowledge element, the "should have known" threshold assessed at the moment of authorization, and not upon the measures element, which remains calibrated to what is "necessary and reasonable" given the authorizing commander's actual capacity at that moment. A commander who possesses, at the moment of authorization, the institutional capacity to require a weapons review and does not do so satisfies both elements simultaneously: constructive knowledge of structural non-compliance arises from failure to conduct the review the circumstances demand; and refusal to require an institutionally available review is a failure to take measures within the commander's actual power. The deployment-stage reorientation proposed here operates entirely within that unaffected standard. A commander who authorizes the deployment of a system that is structurally incapable of generating adequate oversight information breaches this foundational obligation at the very moment of authorization, before any targeting decision is made.[243]

This reorientation is further reinforced by Kreß's confirmation that the crystallization of customary war crimes law in non-international armed conflicts has moved the protective scope of international criminal law to cover all forms of intra-State and transnational armed strife, a

243 Bemba Trial Judgment, *supra* note 233, ¶¶ 197–199; *cf.* Bemba Appeals Judgment, *supra* note 233, ¶¶ 189–193 (reversing conviction upon identifying seven errors of factual assessment in the Trial Chamber's measures analysis-including failure to appreciate limitations faced by a remote commander, failure to assess concrete measures *in concreto*, and erroneous consideration of the accused's motivations-all errors of factual evaluation; the art. 28(1)(a) constructive knowledge standard was neither reformulated nor rejected).

framework within which the deployment authorization of a structurally non-compliant autonomous system squarely falls.[244]

The persuasive authority of the *Bemba Gombo* Trial Chamber Judgment for the present analysis must be assessed in light of the Appeals Chamber's reversal of the conviction on 8 June 2018 (ICC-01/05-01/08-3636-Red), which turned primarily on the Trial Chamber's inadequate factual assessment of *Bemba*'s effective authority and the reasonableness of the measures he took given his position as a remote commander deploying forces to a foreign country, and of the specific measures within his actual capacity in those particular circumstances, not on any rejection of the Article 28(1)(a) treaty standard or its constructive knowledge formulation.[245] The "should have known" obligation accordingly remains a positive and operative treaty standard, confirmed by the text of Article 28(1)(a) Rome Statute itself, and by the ICTY Appeals Chamber's unbroken application of the cognate "had reason to know" standard: in *Blaškić*, the Appeals Chamber confirmed that a superior is criminally responsible only where information was available that would have put him on notice of subordinate offences, rejecting criminal negligence as a basis of liability; in *Strugar*, the Appeals Chamber held that the standard requires information "sufficiently alarming to justify further inquiry," and that a superior's failure to punish known crimes constitutes

244 Claus Kreß, *Transnational Armed Conflicts*, 15 J. Conflict & Sec. L. 245, at 270, 273 (2010) ("The 'Pure Non-International Armed Conflict Model' offers an appropriate legal framework to apply the law on war crimes committed in non-international armed conflicts to the conduct of the members of the non-State party"; "the crystallization of customary war crimes committed in conflicts not of an international character . . . moved the protective scope of international criminal law beyond inter-State incidents to also cover certain forms of intra-State strife").

245 Bemba Appeals Judgment, *supra* note 233, ¶¶ 170–175, 189–193 (reversing conviction on the grounds that the Trial Chamber failed to properly appreciate the limitations Bemba faced as a remote commander deploying forces to a foreign country, ¶¶ 171–173; erred in treating his motivations as determinative of the adequacy of measures taken, ¶¶ 176–179; failed to assess the specific and concrete measures available *in concreto*, ¶ 170; erred in attributing limitations in mandate and results to Bemba personally, ¶¶ 180–181; and failed to assess the actual number of crimes established, ¶¶ 183–184; the art. 28(1)(a) constructive knowledge standard, as formulated by the Pre-Trial Chamber at ¶¶ 428–429 of ICC-01/05-01/08-424 (June 15, 2009), was neither reformulated nor disturbed).

alarming information triggering the duty to prevent future offences.[246] The deployment-stage reorientation advanced in this dissertation does not rest on the Trial Chamber's application of that standard to the Bemba facts; it rests on the anterior and independent proposition that, where a commander authorizes deployment of a system with known behavioral unpredictability and known potential for civilian-affecting cascade effects, the 'circumstances at the time' within the meaning of Article 28(1)(a) Rome Statute include the commander's deployment-moment knowledge of those characteristics, knowledge that is both legally relevant and factually ascertainable at the point of authorization in a manner categorically unavailable at the moment of each subsequent autonomous engagement. The accountability gap identified in this section, therefore, cannot be closed by reformulating the knowledge standard; it requires relocating the entire inquiry to the stage at which the *Bemba Gombo* obligation to create adequate informational conditions for command is actually capable of being discharged.

The actus reus of Article 28(1)(a) Rome Statute requires that the commander "failed to take all necessary and reasonable measures within his or her power to prevent or repress" the commission of the crimes identified.[247] The determination of what constitutes a "necessary and reasonable measure within the commander's power" is context-specific; in the case of conventional ground forces, it encompasses the issuance of orders, the training of subordinates, and the establishment of reporting

246 Rome Statute, *supra* note 233, art. 28(1)(a) ("owing to the circumstances at the time, "should have known" that the forces were committing or about to commit such crimes"); Blaškić Appeals Judgment, *supra* note 233, ¶¶ 62–64; Prosecutor v. Strugar, Case No. IT-01-42-A, ICTY Appeals Chamber, Judgment, ¶¶ 297–301 (July 17, 2008) [hereinafter Strugar Appeals Judgment] (standard requires information "sufficiently alarming to justify further inquiry"; superior's failure to punish known crimes constitutes alarming information triggering duty to prevent future offences).

247 Rome Statute, *supra* note 233, art. 28(1)(b) (requiring that a military commander "failed to take all necessary and reasonable measures within his or her power to prevent or repress their commission or to submit the matter to the competent authorities for investigation and prosecution"). The Appeals Chamber confirmed that the measures element is assessed against the commander's actual material capacity in the specific circumstances existing at the time, reversing the Trial Chamber's conviction. Bemba Appeals Judgment, *supra* note 233, ¶¶ 171, 173, 193–194. The reversal was confined exclusively to the art. 28(1)(b) measures element; the art. 28(1)(a) constructive knowledge standard was neither reformulated nor disturbed. *Id.* ¶ 194 ("one of the elements of command responsibility under article 28(a) of the Statute was not properly established.

procedures. In the context of Tier 3 ACW deployment, the question is what measures were available to the commander at the moment of authorization that could have prevented the autonomous targeting conduct that subsequently produced the criminally relevant harm.

The measure available to the commander at the deployment stage, and the only measure capable of discharging the precautionary obligation under Article 57(2)(a)(i) AP I before the weapon's activation renders it unavailable, is the conduct of a legally adequate Article 36 AP I weapons review. Article 36 AP I requires the High Contracting Party to "determine whether the weapon's employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party."[248] A commander who authorizes the deployment of a Tier 3 ACW system without a completed Article 36 review, or in disregard of a review's finding that the system's structural characteristics produce IHL incompatibility, has failed to take a measure that was: (a) legally required as a treaty obligation under Article 36 AP I for High Contracting Parties, and as a matter of customary obligation derived from the substantive prohibition on indiscriminate weapons for all States; (b) entirely within the commander's institutional authority at the moment of authorization; and (c) the only available mechanism through which IHL compliance of the system's subsequent autonomous targeting conduct could have been assessed before deployment. This analysis does not assert that the ATO's three-component framework constitutes independently established lex lata; it applies the ATO as a matter of *de lege interpretanda* to identify the

248 AP I, art. 36 ("In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party"). The ICRC has confirmed that art. 36 "implies the obligation to establish internal procedures for the purpose of elucidating the issue of legality," that its "aim . . . is to prevent the use of weapons that would violate international law in all circumstances," and that the review obligation "is arguably one that applies to all States, regardless of whether or not they are party to Additional Protocol I." Int'l Comm. of the Red Cross, *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977*, at 5-6, (Kathleen Lawand principal author, rev. Nov. 2006) [hereinafter ICRC Article 36 Guide], https://www.icrc.org/sites/default/files/external/doc/en/assets/files/other/icrc_002_0902.pdf (last visited Apr. 16, 2026); *see also supra* Chapter 2, Section V.3.

specific omission that satisfies the Article 28(1)(a) measures standard at the authorization stage. The Article 36 review obligation is not delegable to the weapon system itself; it is a human legal obligation whose discharge is a precondition of any lawful deployment authorization, and whose omission at that stage constitutes the failure to act that satisfies the actus reus element of Article 28(1)(a) of the Rome Statute's measures obligation, and simultaneously the affirmative omission that may, where Article 25(3)(b) conditions are met, give rise to ordering liability as analyzed in Section I.4 below.

The deployment-stage reorientation of the mens rea standard established in the preceding section and the Article 36 review omission as actus reus are analytically complementary and jointly constitute the full Article 28(1)(a) liability analysis for ACW authorization. A commander who, at the moment of deployment authorization, "should have known", owing to the architectural characteristics of the Tier 3 system, its behavioral opacity, its irreversible propagation capacity, and its structural elimination of real-time human verification, that autonomous engagement decisions would subsequently be taken without IHL-compliant verification of target status, and who nonetheless authorized deployment without completing a weapons review, satisfies both elements of Article 28(1)(a): the mens rea condition (constructive knowledge assessed at authorization) and the actus reus (the failure to conduct the Article 36 review that was the necessary and reasonable measure within the commander's power to prevent the subsequent violations).[249]

This dissertation is structurally distinct from the *Bemba* situation in a legally significant respect. In Bemba, the failure to prevent war crimes

249 The analytical complementarity of constructive knowledge assessed at authorization and art. 36 review omission as *actus reus* constitutes an original doctrinal proposition of this dissertation. Each component rests on independently established authority: the constructive knowledge threshold, Rome Statute, *supra* note 233, art. 28(1)(a); Blaškić Appeals Judgment, *supra* note 242, ¶¶ 62–64; Strugar Appeals Judgment, *supra* note 246, ¶¶ 297–301; the measures standard calibrated to actual material capacity at authorization, Bemba Appeals Judgment, *supra* note 233, ¶¶ 171–173, 193; the art. 36 review as the measure available and legally required at that stage, AP I, art. 36; the ATO framework as the interpretive structure through which the authorization-stage omission is identified, *supra* Chapter 2, Section V.3; Chapter 4, Section I.3.

was assessed against the commander's effective control over forces operating in the field in the Central African Republic, conduct that occurred in a different geographic and operational theater from the commander's location; the Trial Chamber's analysis centered on whether Bemba had the material ability to intervene in ongoing operations.250 In the ACW context, the relevant failure is not assessed against control over post-deployment autonomous targeting conduct, which, as established in Chapter 3, Section II, is precisely the phase over which effective control is structurally absent. It is assessed against control over the development and authorization process itself, which is categorically within a military commander's institutional authority and cannot be delegated to the weapon system whose authorization is under evaluation. The absence of effective control over the weapon's post-activation targeting phase does not preclude liability at the authorization stage; on the analysis advanced in this Chapter, it is precisely because effective control is structurally absent at the operational phase that the accountability inquiry must be relocated to the moment of authorization, where such control indisputably exists.

Applied operationally, this analysis requires that commanders obtain, prior to deployment authorization: technical assessments, from qualified experts, of whether the system's architecture permits adequate monitoring of targeting decisions; characterization of the system's failure modes sufficient to permit advance determination of intervention thresholds; and verification that adequate deactivation mechanisms are functional

250 The Trial Chamber's measures analysis centered on whether Bemba had the material ability to intervene in ongoing operations of the MLC forces in the CAR. Bemba Trial Judgment, *supra* note 233, ¶¶ 405–408 (setting out the command structure, communications capacity, and operational deployment of MLC forces in the CAR as the factual predicate for the measures assessment). The Appeals Chamber reversed on precisely this ground. Bemba Appeals Judgment, *supra* note 233, ¶¶ 171–173. Both the Trial Chamber's analysis and the Appeals Chamber's reversal thus presuppose a commander whose relevant control deficit arose from geographic and operational distance from ongoing field operations-a constraint categorically absent in the ACW authorization scenario, where the authorizing commander acts at the moment of deployment decision with full institutional access to the weapons review apparatus, before any autonomous targeting conduct has occurred.

and operationally tested.[251] Article 87 of AP I reinforces this conclusion: it imposes an affirmative obligation on commanders to prevent breaches and, where necessary, to suppress them and report violations to competent authorities - an obligation that extends, on the analysis advanced in this Section, to structural measures at the design and authorization stage, not merely to reactive measures after violations occur.[252] This doctrinal reorientation engages, and reinforces, the Article 36 weapons-review obligation analyzed in Chapter 2. The Article 36 review is the *ex-ante* mechanism through which the command responsibility accountability threshold - relocated, under the analysis advanced in this Chapter, from the moment of targeting decision to the moment of deployment authorization - is operationalized as a binding legal constraint: a State that authorizes deployment of an autonomous cyber weapon without conducting a weapons review capable of confirming the system's compliance with the ATO's three constitutive components has breached the Article 36 obligation at the moment of authorization, independently of any subsequent targeting decision made by the system.[253]

This reorientation narrows the responsibility gap without closing it entirely, a limitation that intellectual honesty requires acknowledging. A commander held responsible for authorizing an inadequately assessed system has not thereby been held responsible for the specific harm caused; the structural disconnection between human decision and algorithmic act that generates the responsibility gap in the perpetrator model

251 Marco Sassòli, *International Humanitarian Law: Rules, Controversies, and Solutions to Problems Arising in Warfare*, §§ 10.079–10.082 (2d ed., Edward Elgar 2024) [hereinafter Sassòli, IHL] (establishing that meaningful human control at the deployment authorization stage requires, at minimum: awareness of the system's capabilities and limitations (§ 10.079), weaponeering authority over the deployment decision (§ 10.080), and active context controls imposing spatial, temporal, and target-category constraints (§§ 10.081–10.082)-criteria that operationalize the authorization-stage accountability inquiry advanced in this Section and directly support the art. 36 AP I weapons-review obligation analyzed in Chapter 2).

252 AP I, arts. 86(2), 87(1)–(3); ICRC Article 36 Guide, *supra* note 248, at 5–7.

253 AP I, art. 36; ICRC Article 36 Guide, *supra* note 248, at 5–6, 17–18 ("The aim of Article 36 is to prevent the use of weapons that would violate international law in all circumstances . . . by determining their lawfulness before they are developed, acquired or otherwise incorporated into a State's arsenal"; "the review obligation 'implies the obligation to establish internal procedures for the purpose of elucidating the issue of legality'" (quoting 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 1470, 1482)); *see also supra* Chapter 2, Section V.3.

reappears in the command responsibility model in a temporally displaced form. The solution relocates accountability to a moment where it is analytically coherent, the authorization decision, while necessarily conceding that full accountability for specific wrongful acts cannot be reconstructed through doctrinal extension alone. This concession motivates the structural remedies examined in Section II below.

4. Individual Criminal Responsibility under Article 25(3)(b) Rome Statute: Ordering the Deployment of a Tier 3 ACW System

Article 25(3)(b) Rome Statute provides that a person shall be individually criminally responsible if that person "orders, solicits or induces the commission of" a crime within the jurisdiction of the Court "which in fact occurs or is attempted."[254] The mode of liability of ordering within Article 25(3)(b) requires three elements: first, the ordering person must be in a position of authority over the person who executes the criminal act; second, there must be an instruction directed, at least in its foreseeable consequence, at the commission of a crime; and third, a causal relationship must exist between the order and the commission or attempted commission of the crime.[255]

These three elements present a specific configuration when applied to the authorization of Tier 3 ACW deployment that differs from their application in conventional command contexts. A commander who issues an authorization for the deployment of a Tier 3 ACW system is in a position of authority within the meaning of Article 25(3)(b): the

254 Rome Statute, *supra* note 233, art. 25(3)(b) ("Orders, solicits or induces the commission of such a crime which in fact occurs or is attempted"). The Katanga Trial Chamber confirmed that art. 25(3)(b) constitutes a form of accessorial liability, distinct from the principal liability of art. 25(3)(a). Katanga Trial Judgment, *supra* note 417, ¶¶ 1383–1387.

255 Prosecutor v. Mudacumura, Case No. ICC-01/04-01/12, Pre-Trial Chamber II, Decision on the Prosecutor's Application under Article 58, ¶ 63 (Int'l Crim. Ct. July 13, 2012) [hereinafter Mudacumura, *Art. 58 Decision*] (holding that criminal responsibility under art. 25(3)(b) requires that: "(a) the person is in a position of authority; (b) the person instructs another person in any form to either: (i) commit a crime which in fact occurs or is attempted or (ii) perform an act or omission in the execution of which a crime is carried out; (c) the order had a direct effect on the commission or attempted commission of the crime; and (d) the person is at least aware that the crime will be committed in the ordinary course of events as a consequence of the execution or implementation of the order"). The Chamber further confirmed that "the person can give the order through an intermediary and need not give the order directly to the physical perpetrator." *Id.* ¶ 63.

authorization is the operative legal instrument through which the weapon enters active operational status. The authorization constitutes an instruction whose execution, the weapon's subsequent autonomous targeting conduct, may, in the ordinary course of the weapon's designed operation, result in the commission of war crimes, specifically the unlawful attack on civilian objects or the infliction of disproportionate incidental civilian harm under Article 8(2)(b)(ii) and (iv) of the Rome Statute.[256] The causal relationship between the authorization and the crime is established where the deployment authorization is the direct and operative cause of the autonomous targeting decisions that subsequently produce the criminally relevant harm, without any independent human decision-maker intervening between the authorization and the act.

The distinction between Article 28 command responsibility and Article 25(3)(b) ordering liability is both conceptually and operationally significant in the ACW context. Article 28 attaches to the failure to prevent or repress subordinates' crimes, a liability rooted in the omission of oversight. Article 25(3)(b) attaches to the affirmative act of issuing the instruction that produces the crime, a commission-based liability that does not presuppose the existence of a human subordinate executing the order. This distinction matters specifically for Tier 3 ACW, because a commander who cannot be held liable under Article 28(1)(a) due to the structural absence of effective control over the post-deployment autonomous operational phase, the phase during which engagement decisions are taken machine-autonomously without a human chain of command, may nonetheless be liable under Article 25(3)(b) for the authorization decision itself.

The commission of crimes does not need to be certain at the time of authorization; it must be a foreseeable consequence in the ordinary course of the weapon's operation, within the meaning of Article 30(2)(b)

256 Rome Statute, *supra* note 233, arts. 8(2)(b)(ii), 8(2)(b)(iv).

of the Rome Statute.257 Where the four-characteristic framework established in Chapter 1 demonstrates that Tier 3 ACW predictably produce effects on civilian infrastructure through network propagation that no human authorized at the moment of each engagement, the foreseeability standard is satisfied at the authorization stage.

The two modes of liability, Article 28(1)(a) reoriented to the deployment stage, as developed in this Chapter's preceding sections, and Article 25(3)(b) applied to the authorization act, operate cumulatively in the ACW accountability framework, each addressing a distinct dimension of the same underlying decision. Article 28(1)(a) imposes liability for the omission of the Article 36 AP I weapons review that was the necessary and reasonable pre-deployment measure to prevent subsequent violations; Article 25(3)(b) imposes liability for the affirmative act of authorization itself, where the commission of crimes was a foreseeable result. Together, these two provisions ensure that the deployment authorization decision, the only moment at which a human being exercised legally cognizable authority over the weapon's operation, is the moment at which international criminal responsibility attaches, regardless of whether effective control over the post-activation autonomous targeting phase can be established. This cumulative structure responds to the accountability gap identified in Section I of this Chapter: where no individual exercised effective control over the weapon's post-activation conduct, the law does not conclude that no one is responsible; it locates responsibility at the deployment authorization stage and applies the two available modes of liability in parallel.

257 Rome Statute, *supra* note 233, arts. 30(2)(b), 30(3). The Mudacumura Pre-Trial Chamber expressly aligned the art. 25(3)(b) *mens rea* threshold with this standard. Mudacumura, *Art. 58 Decision*, *supra* note 255, ¶ 63.

II. The Accountability Gap: Structural Character and Prospective Remedies

The command responsibility analysis of Section I established that the accountability deficit generated by Tier 3 autonomous cyber weapons is structural, not evidentiary: it cannot be closed by reformulating the knowledge standard, extending the causal chain, or improving investigative methodology, because the deficit arises from the constitutive elimination of the human decision-maker that every existing accountability framework presupposes. That structural finding has a doctrinal consequence that extends beyond IHL and international criminal law. International human rights law — which applies concurrently with IHL during armed conflict as a complementary and non-exclusive body of obligations — generates its own accountability requirements in respect of lethal force, requirements that are cumulative with and not displaced by the command responsibility framework of Article 28 of the Rome Statute. Section II establishes that cumulative accountability dimension before turning, in Sections III and IV, to the structural remedies the deficit demands.

1. The ICCPR Right to Life as a Cumulative Accountability Obligation

The accountability deficit this chapter identifies is not confined to the IHL framework; it operates with equal force under international human rights law, whose obligations remain applicable during armed conflict as a complementary and non-exclusive body of law concurrently with IHL.

Article 6(1) ICCPR states that "every human being has the inherent right to life" and that "no one shall be arbitrarily deprived of his life"; that guarantee is non-derogable, applying without suspension even in situations of armed conflict and other public emergencies that threaten the

life of the nation.[258] The UN Human Rights Committee, the treaty body vested with authoritative interpretive competence over the Covenant, addressed the deployment of autonomous weapons systems directly in General Comment No. 36, stating that "the development of autonomous weapon systems lacking in human compassion and judgment raises difficult legal and ethical questions concerning the right to life, including questions relating to legal responsibility for their use," and concluding that "such weapon systems should not be developed and put into operation, either in times of war or in times of peace, unless it has been established that their use conforms with Article 6 and other relevant norms of international law."[259]

The ICJ confirmed in the *Nuclear Weapons* Advisory Opinion that Article 6 applies concurrently with IHL during hostilities and that what constitutes an arbitrary deprivation of life in that context falls to be determined by the applicable *lex specialis*; the Committee itself states in General Comment No. 36, paragraph 64, that in armed conflict "both spheres of law are complementary, not mutually exclusive."[260]

A commander who authorizes deployment of a Tier 3 autonomous cyber weapon that is structurally incapable of satisfying Article 51(5)(b) and

---

258 International Covenant on Civil and Political Rights art. 6(1), *opened for signature* Dec. 19, 1966, 999 U.N.T.S. 171 (*entered into force* Mar. 23, 1976) [hereinafter ICCPR]; Human Rights Committee, *General Comment No. 36 on Article 6 of the International Covenant on Civil and Political Rights, on the Right to Life*, ¶ 2, U.N. Doc. CCPR/C/GC/36 (Sept. 3, 2019) [hereinafter HRC General Comment No. 36] ("The right to life is the supreme right from which no derogation is permitted, even in situations of armed conflict and other public emergencies that threaten the life of the nation."); *id.* ¶ 67 ("Article 6 is included in the list of non-derogable rights in Article 4(2) of the Covenant. Hence, the guarantees against arbitrary deprivation of life contained in Article 6 continue to apply in all circumstances, including in situations of armed conflict and other public emergencies.").

259 HRC General Comment No. 36, *supra* note 258, ¶ 65 ("The development of autonomous weapon systems lacking in human compassion and judgment raises difficult legal and ethical questions concerning the right to life, including questions relating to legal responsibility for their use. The Committee is therefore of the view that such weapon systems should not be developed and put into operation, either in times of war or in times of peace, unless it has been established that their use conforms with Article 6 and other relevant norms of international law."); *see also* Human Rights Watch & Int'l Human Rights Clinic (IHRC), *A Hazard to Human Rights* 11 (Apr. 2025) (confirming that the Committee "specifically addressed the development and use of autonomous weapons systems in General Comment No. 36" and that such systems "should not be developed and put into operation, either in times of war or in times of peace," unless conformity with art. 6 is positively established).

260 Nuclear Weapons Advisory Opinion, *supra* note 39, ¶ 25 ("The test of what is an arbitrary deprivation of life, however, then falls to be determined by the applicable *lex specialis*."); HRC General Comment No. 36, *supra* note 258, ¶ 64 ("Article 6 continues to apply also in situations of armed conflict to which the rules of international humanitarian law are applicable. . . . Both spheres of law are complementary, not mutually exclusive.").

57(2)(a)(i) AP I simultaneously authorizes an operation incapable of conforming with Article 6(1) ICCPR; the accountability obligations generated by the right to life are accordingly cumulative with, and not displaced by, the command responsibility framework of Article 28 of the Rome Statute.[261]

2. Structural Determinants of the Accountability Deficit

The analysis in this Chapter — Sections I and II — reveals a consistent pattern, at each level of the international accountability architecture, State attribution under ARSIWA, individual criminal liability under the Rome Statute, and command responsibility under Article 86 AP I, the foundational presuppositions of the framework are defeated by the same feature of autonomous cyber weapons: the replacement of contemporaneous human decision–making by algorithmically generated choices across a distributed, temporally extended causal chain. This accountability deficit is structural in the technical sense: it arises not from evidentiary inadequacy or interpretive limitation but from a categorical mismatch between the conceptual architecture of accountability law and the operational architecture of autonomous systems. It cannot be closed by greater investigative diligence, improved prosecutorial strategy, or creative interpretation of existing doctrine.

Sparrow's 2007 analysis provides the clearest statement of the problem's logical structure: as autonomous systems grow more sophisticated, the human control that accountability frameworks presuppose becomes progressively less achievable, until we confront the prospect of sophisticated machines capable of causing serious harm without any clearly identifiable individual or institution bearing responsibility for the deaths that result.[262] This analysis was directed at kinetic autonomous weapons;

---

261 Human Rights Committee, *General Comment No. 31 on the Nature of the General Legal Obligation Imposed on States Parties to the Covenant*, ¶ 11, U.N. Doc. CCPR/C/21/Rev.1/Add.13 (May 26, 2004) [hereinafter HRC General Comment No. 31]; *see also* Rome Statute, *supra* note 233, art. 28(1)(a).

262 Robert Sparrow, Killer Robots, 24 J. Applied Phil. 62, at 70-72 (2007) [hereinafter Sparrow, Killer Robots], https://robsparrow.com/wp-content/uploads/Killer-robots.pdf (last visited Apr. 16, 2026); S.G. Report, *LAWS 2024*, *supra* note 9, ¶ 32.

it applies with compounded force to autonomous cyber weapons operating at machine speed across complex network environments, where the temporal and causal distance between human decision and machine act is greatest, and where the opacity of machine–learning architectures is compounded by the intrinsic attribution difficulties of cyber operations documented in Chapter 2.[263]

This chapter advances the further claim, beyond Sparrow's diagnostic observation, that the structural responsibility gap is not merely a technical legal problem but a structural defect in the deterrence architecture upon which IHL's civilian-protection function depends. The accountability frameworks of IHL and international criminal law derive their protective function primarily through deterrence: the prospect of individual and State responsibility motivates precautionary behavior by decision-makers before harm occurs. If no individual and no State can be held responsible for harm caused by an autonomous system in a manner that satisfies the legal framework's foundational requirements, the accountability mechanism through which IHL exerts its normative influence on the conduct of warfare is severed. The law's protective purpose - preventing civilian harm - is correspondingly compromised not merely at the level of post hoc remedy but at the level of the incentive structure that shapes *ex-ante* conduct.[264] This normative urgency establishes that the governance obligations are legally necessary rather than merely desirable: governance reform is required not only to improve the practical effectiveness of IHL but to preserve its normative function in conditions where accountability frameworks have been rendered structurally inoperative.

263 Tallinn Manual 2.0, *supra* note 18, r. 14 cmt. ¶¶ 1, 4–5, r. 15 cmt. ¶¶ 13–16; Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security, *Report of the Group of Governmental Experts*, ¶¶ 14, 21–22, U.N. Doc. A/76/135 (July 14, 2021), https://documents.un.org/doc/undoc/gen/n21/166/74/pdf/n2116674.pdf (last visited Apr. 16, 2026) [hereinafter 2021 GGE Cyber Report].

264 Sparrow, *Killer Robots*, *supra* note 262, at 67–68; Rome Statute, *supra* note 233, pmbl. ¶ 5; Boulanin & Bo, *Three Lessons*, *supra* note 264.

3. Structural Remedies: Treaty Revision, Institutional Design, and Normative Attribution

Four structural remedies have been advanced in scholarship and in intergovernmental negotiations under the CCW,[265] which employs the broader operative category of lethal autonomous weapons systems ("LAWS"), within which autonomous cyber weapons constitute a subcategory. The first is strict State liability for the deployment of autonomous cyber weapons. The 1972 Convention on International Liability for Damage Caused by Space Objects provides the most developed international precedent, imposing absolute liability on launching States for damage caused by space objects on the surface of the Earth, irrespective of fault. The analogy is limited: Article I(a) of the Liability Convention defines compensable damage as physical harm to persons or property, whereas the predominant harm caused by autonomous cyber weapons, service disruption, data destruction, and network degradation is non-physical and therefore falls outside the Convention's damage definition as a matter of its express terms. The analogy accordingly requires treaty modification extending the liability regime to non-physical harm, or the development of an equivalent regime *de lege ferenda*.[266] Applied analogically to autonomous cyber weapons, a strict liability regime would make the causal connection between deployment and harm, rather than direction of the specific wrongful act, the operative basis for State accountability, thereby circumventing the attribution requirement under ARSIWA by substituting causal nexus between deployment and harm for the direction-or-control threshold of Articles 8 and 11 ARSIWA. This approach also creates powerful pre–deployment

265 CCW GGE, *Rolling Text*, U.N. Doc. CCW/GGE.1/2024/WP.X (2024), ¶¶ II.5, III.1, IV.1, V.1 [hereinafter CCW GGE, *Rolling Text (2024)*]; ICRC Article 36 Guide, *supra* note 248.

266 Convention on International Liability for Damage Caused by Space Objects art. II, Mar. 29, 1972, 961 U.N.T.S. 187 [hereinafter Space Liability Convention], https://www.unoosa.org/oosa/en/ourwork/spacelaw/treaties/liability-convention.html (last visited Apr. 16, 2026); Alisha Anand & Ioana Puscas, *Proposals Related to Emerging Technologies in the Area of Lethal Autonomous Weapons Systems: A Resource Paper* 3, 28–29 (UNIDIR Sec. & Tech. Programme 2023), https://unidir.org/files/2023-05/UNIDIR-Proposals_Emerging_Technologies_Lethal_Autonomous_Weapons_Systems_2023.pdf (last visited Apr. 16, 2026) [hereinafter Anand & Puscas, *UNIDIR Resource Paper*].

incentives: States bearing absolute financial liability and the reputational and legal consequences of inter-State claims proceedings for autonomous system failures have compelling reasons to ensure IHL compliance. Its limitation is that it operates exclusively at the inter–State level and cannot supply individual criminal accountability for egregious or systematic violations. It also encounters serious evidentiary challenges in the cyber context, where the prior problem of technical attribution, establishing that a cyber operation originated from a specific State-deployed autonomous system rather than a third-party actor or the State itself, must be resolved before the Liability Convention's causal framework can operate, a threshold challenge the Convention's physical damage paradigm was not designed to address.[267]

The second remedy is expansion of command responsibility to encompass systemic risk liability: as a matter of *de lege ferenda*, commanders should bear criminal liability for creating systemic conditions of harm through deployment of inadequately tested or constitutively opaque systems, building on the doctrinal reorientation proposed in Section I.3 of this Chapter. This approach is adoptable by existing institutions without treaty amendment and builds on established jurisprudence. Its limitation is that it reaches commanders at the authorization stage while leaving the broader decision chain, designers, manufacturers, and political authorities who commission weapons acquisition, substantially unaddressed. The third remedy, corporate manufacturer liability under international law, draws on the UN Guiding Principles on Business and Human Rights, which articulate, as a matter of soft law, that business enterprises bear a responsibility to respect human rights and to provide

267 ARSIWA, *supra* note 21, arts. 8, 58; Laura Bruun, *Towards a Two-Tiered Approach to Regulation of Autonomous Weapon Systems: Identifying Pathways and Possible Elements* v, 2, 5 (SIPRI, Aug. 2024), https://www.sipri.org/sites/default/files/2024-08/laws.pdf (last visited Apr. 16, 2026) [hereinafter Bruun, *Two-Tiered Approach*].

remedies for violations attributable to their products.[268] Applied to autonomous weapons manufacturers, this framework would impose responsibility for design–level failures of IHL compatibility. The analytical challenge is that the Guiding Principles currently express a non–legally binding corporate responsibility rather than a judicially enforceable legal obligation; incorporating them into a binding accountability regime would require either treaty codification or identification of a customary rule of equivalent effect, neither of which is currently available.

The fourth remedy, establishment of a specialized international tribunal with jurisdiction over autonomous weapons violations, represents the most institutionally comprehensive response. Such a body could develop hybrid attribution standards combining State and individual accountability, establish evidentiary standards adapted to the distinctive challenges of autonomous systems and cyber operations, and provide the adjudicative capacity that existing courts, designed for human perpetrators operating in recognizable command hierarchies, currently lack.[269] The political feasibility of this option in the present geopolitical environment is, however, limited: the consensus on the value of multilateral institutional development that enabled the creation of the ICTY, ICTR, and ICC does not characterize the current international moment,

268 U.N. Office of the High Commissioner for Human Rights, *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework*, princs. 11, 13(b), 22, U.N. Doc. HR/PUB/11/04 (2011), *endorsed by* Human Rights Council Res. 17/4, U.N. Doc. A/HRC/RES/17/4 (June 16, 2011), https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf (last visited Apr. 16, 2026) [hereinafter UNGP on Business and Human Rights].

269 G.A. Res. 78/241, ¶¶ 1–2, U.N. Doc. A/RES/78/241 (Dec. 22, 2023), https://docs.un.org/en/a/res/78/241 (last visited Apr. 16, 2026); G.A. Res. 79/62, *Lethal Autonomous Weapons Systems*, ¶¶ 3–5, U.N. Doc. A/RES/79/62 (Dec. 2, 2024), https://docs.un.org/en/a/res/79/62 (last visited Apr. 16, 2026); ICRC Article 36 Guide, *supra* note 248, at 2–4, 5–7, 20–22.

and the major military powers that would need to accept jurisdiction have shown no disposition to do so.[270]

4. Synthesis: The Ex–Ante Governance Imperative

Each of the four structural remedies offers partial but individually insufficient relief. Strict State liability resolves the attribution problem at the inter–state level but does not supply individual criminal accountability. Expanded command responsibility applies to a limited class of decision-makers at the authorization stage, while leaving the design and manufacturing chain unaddressed. Corporate liability under the Guiding Principles is currently non-binding. A specialized tribunal is conceptually comprehensive but politically unrealizable in the medium term. Even in combination, the four remedies would substantially reduce but not eliminate the structural accountability deficit: the categorical mismatch between autonomous decision-making and accountability law's foundational presuppositions is sufficiently deep that it cannot be

270 The absence of multilateral consensus on a legally binding instrument is demonstrated by the record of the CCW Group of Governmental Experts on LAWS, within which delegations-including major military powers-have persistently diverged on the scope of characterization, the nature of any possible instrument, and the content of binding prohibitions and regulations. *See* CCW GGE, *Chair's Summary-First 2025 Session of the GGE on LAWS*, CCW/GGE.1/2025/WP.1, ¶¶ 11–14, 21–24 (Apr. 7, 2025) (recording persistent divergences on lethality, scope, and the prescriptive character of proposed regulations, with several delegations opposing binding measures as exceeding the Group's mandate or as premature given technological uncertainty). *See also* ICRC, *AWS Selected Issues*, *supra* note 15, at 1, 22 (noting that "without timely agreement on a legally binding instrument that prohibits unacceptable AWS and regulates all others, emerging patterns in the research, development and use of AWS may become difficult to reverse, with lasting humanitarian and legal consequences," and characterizing adoption of binding rules as "an urgent humanitarian and legal priority"). The absence of major-power acceptance of expanded international criminal jurisdiction is demonstrated by the non-participation of the United States, the Russian Federation, and the People's Republic of China in the Rome Statute system. The United States withdrew its signature by letter to the U.N. Secretary-General of May 6, 2002; *see* John R. Bolton, *Letter to Kofi Annan, Secretary-General of the United Nations* (May 6, 2002), *reprinted in* 41 I.L.M. 1014 (2002). The Russian Federation withdrew its signature by Presidential Bylaw № 361-rp of November 16, 2016 (*O namerenii Rossiyskoy Federatsii ne stat' uchastnitsey Rimskogo statuta Mezhdunarodnogo ugolovnogo suda* [On the Russian Federation's Intention Not to Become a Party to the Rome Statute of the International Criminal Court]), publication.pravo.gov.ru, Doc. No. 0001201611160018; *see also* U.N. Treaty Collection, *Multilateral Treaties Deposited with the Secretary-General*, ch. XVIII, § 10, n. 9 (recording receipt by the Secretary-General on November 30, 2016 of the Russian Federation's notification pursuant to Vienna Convention on the Law of Treaties, *supra* note 89, art. 18(a)), https://treaties.un.org/pages/ViewDetails.aspx?src=TREATY&mtdsg_no=XVIII-10&chapter=18 (status as at June 20, 2026). The People's Republic of China has not signed the Rome Statute. *Id. See also* Claus Kreß & Leonie von Holtzendorff, *The Kampala Compromise on the Crime of Aggression*, 8 J. Int'l Crim. Just. 1179, 1208–1212 (2010) (analyzing how the United States, as a non-State party, and permanent members of the Security Council resisted any jurisdictional regime that would allow the ICC to exercise jurisdiction over the crime of aggression without Security Council authorization, tracing the formidable structural resistance to expanded ICC jurisdiction from major military powers as a persistent feature of negotiations on the crime of aggression).

eliminated through institutional innovation operating exclusively at the post–harm end of the legal system.

The conclusion to which this analysis drives is therefore not that existing law is incapable of incremental improvement — the due diligence framework, the extended command responsibility doctrine, and analogical application of the Liability Convention's strict-liability model (subject to the treaty-modification requirement identified in Section II.3 above) each have genuine, if partial, analytical traction — but that these ex post accountability mechanisms, however creatively developed, are insufficient as a matter of principle. What is additionally and necessarily required is a set of ex ante governance obligations operating at the design, review, and authorization stages — obligations that prevent the deployment of systems whose architecture makes IHL compliance and meaningful oversight structurally impossible before harm occurs. This is the analytic bridge to Chapter 5: the governance obligations developed in that Chapter — design requirements mandating interpretable outputs, mandatory Article 36 review protocols incorporating machine learning failure-mode assessment, temporal re-verification obligations throughout the deployment lifecycle, and certification regimes establishing minimum IHL-compliance standards— represent not policy desiderata but legal necessities, grounded in the protective purpose of IHL itself and in the normative imperative established in this Chapter.[271]

The three contributions of this Chapter converge on a single finding whose implications extend beyond the specific doctrinal questions addressed: autonomous cyber weapons do not merely complicate the application of existing accountability doctrine but structurally defeat the normative architecture upon which that doctrine's deterrence function depends. A law of command responsibility that cannot identify the responsible individual, applied to a weapons architecture designed to

271 AP I, art. 36; Nuclear Weapons Advisory Opinion, *supra* note 39, ¶ 78.

eliminate that individual, does not produce a gap in the law — a lacuna remediable by analogy under Article 38(1)(c) ICJ Statute. It produces a structural void: a categorical mismatch between accountability law's conceptual architecture and the operational architecture of autonomous systems that no interpretive technique can resolve. The governance response developed in Chapter 5 is the institutional consequence that the finding requires.

Chapter 5

The Human Security by Design Approach

Chapters 2 and 3 established two findings.

First: IHL's precautionary architecture cannot operate for Tier 3 autonomous cyber weapons because those weapons eliminate the human decision-maker IHL presupposes.

Second: accountability law cannot operate at the moment of autonomous targeting for the same reason — no human is present.

Both findings point to the same place: the only moment in the entire lifecycle of an autonomous cyber weapon where a human decision-maker still exists is the design phase. That is where the law must act.

This Chapter translates IHL's existing treaty obligations into four enforceable technical design criteria that a system must satisfy before it can be lawfully deployed. These are not new law.

They are what IHL already requires, stated in language engineers can build and lawyers can verify:

- a kill-switch override (Article 57(2)(b) AP I),
- geofencing constraints (Article 48 AP I),
- temporal deactivation logic (Article 57(1) AP I),
- and tamper-proof audit logging (Article 28(1) Rome Statute).

A system that satisfies all four has been redesigned so as to no longer eliminate the human at the operative moment — it has, in doctrinal terms, been reclassified from Tier 3 to Tier 2.
A system that cannot satisfy them cannot be lawfully deployed under existing law. Human Security by Design is the name for this framework. It is not a new legal regime. It is the implementation architecture that existing law demands.

I. Human Security as a Normative Framework: Definitional Contours and Digital-Age Application

The following table operationalizes those protection frameworks across all seven human security dimensions.

Table 1: Human Security Components, Critical Infrastructure Vulnerabilities, and Legal Protection Framework

Human Security Dimension	Critical Infrastructure Systems	Autonomous Cyber Weapon Threats	Applicable Legal Protection Framework
Economic Security	• Financial networks, banking systems, stock exchanges • Payment platforms, cryptocurrency exchanges • Central bank digital currency infrastructure	• Ransomware targeting financial transaction servers • Denial–of–service attacks disrupting banking infrastructure • Algorithmic manipulation of trading systems causing market destabilization	• **AP I Articles 48, 52(1), 52(2), 54(2)** (distinction; definition of military objectives; protection of objects indispensable to civilian survival) • **Customary IHL Rules 7–10**[272]

272 Jean-Marie Henckaerts & Louise Doswald-Beck, Customary International Humanitarian Law rr. 7–10 (Cambridge Univ. Press 2005.

Table 1 (continued)

Human Security Dimension	Critical Infrastructure Systems	Autonomous Cyber Weapon Threats	Applicable Legal Protection Framework
Health Security	• Hospital information systems and networked medical devices • Pharmaceutical supply chains and vaccine distribution networks • Public health surveillance and telemedicine platforms	• Ransomware targeting hospital networks (WannaCry, May 2017;[273] Ryuk, September 2020)[274] • Manipulation of medical device parameters (insulin pumps, pacemakers, ventilators) • Compromise of vaccine cold–chain monitoring systems	• **AP I Articles 12–15, 18** *(protection of medical units, personnel, and transports)* • **GC IV Article 18** *(protection of civilian hospitals)* • **Customary IHL Rules 25–30**

273 CISA/NCCIC & FBI, *Alert TA17-132A: Indicators Associated with WannaCry Ransomware* (May 12, 2017, updated June 7, 2018), https://www.cisa.gov/news-events/alerts/2017/05/12/indicators-associated-wannacry-ransomware (last visited Apr. 16, 2026) (tens of thousands of infections in over 150 countries; attributed to North Korean government) [hereinafter CISA, *WannaCry Alert*]; National Audit Office (U.K.), *Investigation: WannaCry Cyber Attack and the NHS* 5 (Oct. 27, 2017), https://www.nao.org.uk/wp-content/uploads/2017/10/Investigation-WannaCry-cyber-attack-and-the-NHS.pdf (last visited Apr. 16, 2026) (80 of 236 NHS trusts affected; 34 locked out of devices; NHS England declared a major incident) [hereinafter NAO, *WannaCry Report*].

274 CISA, FBI & HHS, *Joint Cybersecurity Advisory AA20-302A: Ransomware Activity Targeting the Healthcare and Public Health Sector* 1 (updated Oct. 29, 2020), https://www.cisa.gov/news-events/cybersecurity-advisories/aa20-302a (last visited Apr. 16, 2026) (increased and imminent cybercrime threat to U.S. hospitals and healthcare providers; Ryuk and Conti ransomware deployed via TrickBot and BazarLoader causing disruption of healthcare services) [hereinafter CISA, *Healthcare Ransomware Advisory*].

Table 1 (continued)

Human Security Dimension	Critical Infrastructure Systems	Autonomous Cyber Weapon Threats	Applicable Legal Protection Framework
Personal Security	• Emergency response networks and dispatch systems • Law enforcement and forensic databases • Domestic violence protection and missing persons registries	• Compromise of emergency communication networks preventing dispatch • Manipulation of criminal investigation databases • Ransomware disabling disaster–response coordination platforms	• **Customary IHL Rules 15–21** *(precautionary obligations in attack)* • **ICCPR Article 6(1)** *(inherent right to life)* • **AP I Article 51(2)** *(prohibition of attacks against the civilian population)*

The same structural incompatibility analysis applies to the remaining four human security dimensions; the three examples above establish the framework's application.

Sources: UNDP Human Development Report 1994, at 22–25; G.A. Res. 66/290 (2012); AP I Articles 12–15, 18, 35(3), 48–55; ICRC Customary IHL Study, supra note 75, rr. 7–10, 15–21, 25–30, 45, 54; supra notes 284,287.

Legal protection entries state only *lex lata* rules except where marked "*de lege ferenda* only" in the relevant cell; the operational application of lex lata rules to autonomous cyber weapons via the technical specifications of Sections II–V constitutes the *de lege ferenda* contribution of this Chapter.[275]

Table 1 maps each of the seven components onto its corresponding legal protection architecture. The following paragraphs supply the empirical foundation for that mapping through documented case studies.

Autonomous cyber operations against health infrastructure hospitals, pharmaceutical supply chains, medical devices, public health surveillance systems, and telemedicine platforms directly imperil health security by denying populations access to emergency medical care and disrupting disease surveillance essential for epidemic response.[276] The 2017 WannaCry ransomware incident provided empirical demonstration when the United Kingdom's National Health Service experienced forced cancellation of an estimated 19,494 appointments including operations, with at least 80 out of 236 trusts across England affected or disrupted, of which 34 were directly infected with the ransomware and a further 46 were indirectly disrupted,[277] while the Department of Health and Social Care subsequently estimated total costs at £92 million, and

275 AP I, arts. 12–15, 18, 35(3), 48, 51(2), 52, 53, 54, 55; Geneva Convention Relative to the Protection of Civilian Persons in Time of War art. 18, Aug. 12, 1949, 75 U.N.T.S. 287 [hereinafter GC IV]; Convention for the Protection of Cultural Property in the Event of Armed Conflict arts. 4, 16, May 14, 1954, 249 U.N.T.S. 240 [hereinafter Hague Cultural Property Convention]; Second Protocol to the Hague Convention of 1954 for the Protection of Cultural Property in the Event of Armed Conflict art. 6, Mar. 26, 1999, 2253 U.N.T.S. 172 [hereinafter Second Hague Protocol]; ENMOD Convention, *supra* note 275, arts. I–II; ICCPR, *supra* note 258, arts. 6(1), 25; G.A. Res. 61/295, arts. 11–13, U.N. Doc. A/RES/61/295 (Sept. 13, 2007); Jean-Marie Henckaerts & Louise Doswald-Beck, Customary International Humanitarian Law rr. 7–10, 15–21, 25–30, 45, 54 (Cambridge Univ. Press 2005); Nicaragua, *supra* note 35, ¶¶ 202–209; U.N. Charter art. 2(4), *signed* June 26, 1945, 59 Stat. 1031, T.S. No. 993 (*entered into force* Oct. 24, 1945).

276 Int'l Comm. of the Red Cross, *Protection of Medical Personnel, Units and Transports during Armed Conflict* 1–2 (Feb. 2025), https://www.icrc.org/sites/default/files/2025-02/06_Medical_facilities.pdf (last visited Apr. 16, 2026) [hereinafter ICRC, *Medical Personnel Protection*]; World Health Org., *Cyber-Attacks on Critical Health Infrastructure* (Feb. 6, 2024), https://www.who.int/news-room/questions-and-answers/item/cyber-attacks-on-critical-health-infrastructure (last visited Apr. 16, 2026). While these sources document general cyber operations rather than specifically autonomous systems, the documented impacts-emergency care denial through ambulance diversions, disrupted epidemic response capabilities, and compromised pharmaceutical supply chains-demonstrate the vulnerabilities that autonomous cyber weapons with algorithmic target selection, self-propagation capabilities, and temporally delayed effects would exploit and amplify. *See supra* Chapter 1, Section II (documenting autonomous cyber weapon characteristics).

277 NAO, *WannaCry Report*, *supra* note 273, at 6–7, ¶ 5 & fig. 1.

the official investigation concluded that all infected organizations had shared the same vulnerability and could have taken "relatively simple action to protect themselves" against the attack.[278]

WannaCry, a criminal ransomware campaign rather than an autonomous cyber weapon, is invoked here only as an illustrative precedent: it demonstrates, with empirical precision, the scale of cascading civilian harm a self-propagating algorithm can inflict on a critical civilian sector within hours, precisely the risk profile Article 36's pre-deployment review obligation is designed to forestall. A deliberate autonomous cyber weapon incorporating WannaCry's rapid self–propagation logic deployed during armed conflict would violate AP I Article 48's general distinction principle requiring parties to "at all times distinguish between the civilian population and combatants and between civilian objects and military objectives" and the specific protections Article 12 and 18 afford medical units, rendering the attack indiscriminate under Article 51(4)(c), which prohibits attacks 'which employ a method or means of combat the effects of which cannot be limited as required by this Protocol.'[279]

Food and economic security face structurally comparable threats from autonomous operations affecting agricultural infrastructure, food logistics systems, financial networks, and industrial control systems (ICS) governing manufacturing facilities, with successful operations potentially triggering cascading agricultural production failures, financial

278 NAO, *WannaCry Report*, *supra* note 273, at 6, 14; U.K. Dep't of Health & Social Care, *Securing Cyber Resilience in Health and Care: Progress Update October 2018* 14 (Oct. 2018), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/747464/securing-cyber-resilience-in-health-and-care-september-2018-update.pdf (last visited Apr. 16, 2026) [hereinafter DHSC, *Cyber Resilience Update 2018*].

279 AP I, arts. 48, 12(1), 18(1), 51(4)(c); ICRC, *Distinction Short Paper* [hereinafter ICRC, *Distinction Short Paper*], at 2.

system instability, and industrial production halts.[280] The December 2016 CRASHOVERRIDE (Industroyer) deployment against a single Ukrenergo distribution substation in Kyiv demonstrated proof-of-concept for ICS-targeted malware capable of directly manipulating circuit breakers through four ICS protocol payloads (IEC 60870-5-101, IEC 60870-5-104, IEC 61850, and OPC DA), combined with a wiper component rendering systems unrecoverable and a denial-of-service module targeting protective relay firmware; ESET researchers characterized Industroyer as the "biggest threat to industrial control systems since Stuxnet," possessing modular architecture enabling rapid adaptation to target power grids globally.[281] The May 2021 Colonial Pipeline attack demonstrated parallel vulnerabilities: DarkSide ransomware compromised IT systems, prompting a precautionary shutdown that halted a pipeline transporting 2.5 million barrels daily, approximately 45% of the East Coast's gasoline, diesel, and jet fuel supply, for approximately five days.[282] The disruption generated compounding effects across economic security (fuel shortages, price spikes, and emergency regulatory waivers), personal security (supply disruptions affecting 50 million Americans), and political security (presidential emergency intervention and Department of Energy activation), illustrating the cross-sector cascade

280 Ajay Kulkarni et al., *A Review of Cybersecurity Incidents in the Food and Agriculture Sector* 1–2, 9–10 (Mar. 12, 2024), https://arxiv.org/pdf/2403.08036.pdf (last visited Apr. 16, 2026) [hereinafter Kulkarni et al., *Cybersecurity in Food Sector*]; Int'l Monetary Fund, *Global Financial Stability Report: The Last Mile: Financial Vulnerabilities and Risks* 79, 82–83 (Apr. 2024), https://www.imf.org/en/Publications/GFSR/Issues/2024/04/16/global-financial-stability-report-april-2024 (last visited Apr. 16, 2026) [hereinafter IMF, *GFSR 2024*]; Int'l Comm. of the Red Cross, *International Humanitarian Law and Cyber Operations During Armed Conflicts* 8 (Nov. 2019), https://www.icrc.org/en/document/international-humanitarian-law-and-cyber-operations-during-armed-conflicts (last visited Apr. 16, 2026) [hereinafter ICRC, *IHL & Cyber Ops*].

281 Cherepanov, *supra* note 4, at 5, 13, 15; Dragos, *CRASHOVERRIDE* , *supra* note 4, 3–4, 12.

282 CISA & FBI, *Joint Cybersecurity Advisory AA21-131A, DarkSide Ransomware: Best Practices for Preventing Business Disruption from Ransomware Attacks* (May 11, 2021, last rev. July 8, 2021) [hereinafter CISA, *DarkSide Advisory*] (confirming deployment of DarkSide ransomware against the pipeline company's IT network and noting that the company "proactively disconnected certain OT systems to ensure the systems' safety," with no indication that OT networks were directly affected by the ransomware); *Colonial Pipeline Restarts Operations Days After Hack*, Associated Press, L.A. Times (May 12, 2021) (reporting that Colonial Pipeline "delivers about 45% of the fuel consumed on the East Coast" and restarted operations on Wednesday, May 12, 2021, five days after the May 7, 2021 shutdown, with the company stating that "all lines . . . will return to normal operations" but that "it will take several days for the product delivery supply chain to return to normal"); *Colonial Announces Pipeline Restart, Says Normal Service Will Take 'Several Days'*, NBC News (May 12, 2021) (confirming restart "at approximately 5 p.m. ET" on May 12, 2021 following five-day shutdown caused by DarkSide ransomware).

propagation architecture whose weaponized equivalent, incorporated into an autonomous cyber weapon deployed in armed conflict, would directly engage IHL precautionary obligations under Articles 51(5)(b) and 57 AP I.[283]

Environmental, personal, community, and political security, four of the seven components identified in the United Nations human security framework, face equally grave jeopardy when autonomous cyber weapons target critical civilian infrastructure systems essential to survival, livelihood, and dignity.[284] Water treatment systems demonstrate environmental security vulnerabilities when unauthorized cyber actors manipulate SCADA controls governing chemical dosing parameters, as occurred in February 2021, when unidentified actors accessed the Bruce T. Haddock Water Treatment Plant's SCADA system in Oldsmar, Pinellas County, Florida, and increased the sodium hydroxide concentration from 100 parts per million to 11,100 parts per million before an operator observing the remote access in real time reversed the change, preventing public harm.[285] Electoral infrastructure presents parallel political security threats, documented in the Senate Intelligence Committee's comprehensive investigation concluding that Russian government operations directed "extensive activity against U.S. election infrastructure" beginning in at least 2014, exploiting cybersecurity vulnerabilities in State

283 *Testimony of Joseph Blount, President and Chief Executive Officer, Colonial Pipeline Company, Before the U.S. Senate Committee on Homeland Security and Governmental Affairs* 1–4 (June 8, 2021), https://www.hsgac.senate.gov/wp-content/uploads/imo/media/doc/Testimony-Blount-2021-06-08.pdf (last visited Apr. 16, 2026) [hereinafter Blount, *Senate Testimony*]; U.S. Dep't of Energy, *Colonial Pipeline Cyber Incident* (May 7, 2021), https://www.energy.gov/ceser/colonial-pipeline-cyber-incident (last visited Apr. 16, 2026); White House, *Fact Sheet: The Biden-Harris Administration Has Launched an All-of-Government Effort to Address Colonial Pipeline Incident* (May 11, 2021), https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2021/05/11/fact-sheet-the-biden-harris-administration-has-launched-an-all-of-government-effort-to-address-colonial-pipeline-incident/ (last visited Apr. 16, 2026) [hereinafter White House, *Colonial Pipeline Fact Sheet*].

284 U.N. Development Programme, Human Development Report 1994: New Dimensions of Human Security ch. 2, at 7 (Oxford Univ. Press 1994) [hereinafter UNDP, *Human Development Report 1994*].

285 FBI/CISA, *Water Treatment Alert*, *supra* note 273.

and local systems while developing capabilities designed to undermine confidence in U.S. democratic institutions and voting processes.[286]

The United Nations' seven–dimensional human security framework economic, food, health, environmental, personal, community, and political security establishes these components as "interrelated and mutually reinforcing," requiring "comprehensive" responses to "interlinkages" between security threats as recognized in General Assembly Resolution 66/290,[287] suggesting autonomous cyber weapons affecting critical infrastructure pose systemic threats to this entire architecture rather than isolated, sector–specific risks amenable to narrow technical fixes.

AP I's precautionary framework, elaborated for kinetic warfare against targets with deterministic and spatially contained effects, does not address the cascading, temporally diffuse, and jurisdictionally dispersed harms that autonomous cyber operations against interconnected digital–physical infrastructure systems produce, a structural gap that renders ex post attribution, State responsibility litigation, and individual criminal accountability insufficient as the sole governance response, and that the human security framework addresses by supplying the normative foundation for the ex ante protective architecture this Chapter develops.

States' due diligence obligations under IHL must therefore encompass proactive critical infrastructure protection measures beyond passive compliance with distinction and proportionality requirements assessed retroactively. Article 36 of AP I requires High Contracting Parties to determine "whether the employment of a new weapon, means or method of warfare would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party," creating an affirmative legal obligation to assess

286 S. Select Comm. on Intelligence, *Report on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election*, S. Rep. No. 116-290, vol. 1, at 3, 5 (2019), https://www.intelligence.senate.gov/sites/default/files/documents/report_volume1.pdf (last visited Apr. 16, 2026) [hereinafter SSCI, *Russian Active Measures Report*].

287 UNDP, *Human Development Report 1994*, *supra* note 284, at 7, 15, 17; G.A. Res. 66/290, ¶ 3, U.N. Doc. A/RES/66/290 (Sept. 10, 2012).

autonomous cyber weapons' structural capacity for IHL compliance before operational deployment; Article 57(1) imposes a duty of constant care to spare the civilian population in the conduct of military operations,

Article 57(2)(a)(ii) requires taking 'all feasible precautions in the choice of means and methods of attack with a view to avoiding, and in any event to minimizing, incidental loss of civilian life,' and Article 58 imposes complementary duties on the defending party, to the maximum extent feasible, to protect civilian populations under its control against the effects of attacks - obligations that, in the cyber domain, this dissertation argues must be discharged through embedding technical safeguards directly into weapon systems' architecture and implementing resilient infrastructure design separating civilian from military networks, as analyzed in Section II.2 below.[288]

The ICJ's Advisory Opinion on the Legality of the Threat or Use of Nuclear Weapons confirms that environmental protection is a legally cognizant element in weapons-legality and proportionality assessments, affirming that States must take relevant environmental factors into account in the context of the implementation of the principles and rules of the law applicable in armed conflict.[289] The Court further affirmed that the cardinal principles contained in the texts constituting the fabric of humanitarian law, the protection of civilians through the principle of distinction, and the prohibition of causing unnecessary suffering to combatants, constitute intransgressible principles of international customary law,[290] and that, in application of the latter, States "do not have unlimited freedom of choice of means in the weapons they use."[291] Accordingly,

288 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 1466, 1469–1472, 1478–1482.

289 Nuclear Weapons Advisory Opinion, *supra* note 39, ¶ 30 (holding that, while existing international law does not specifically prohibit the use of nuclear weapons, it indicates important environmental factors that are properly to be taken into account in the context of the implementation of the principles and rules of the law applicable in armed conflict).

290 *Id.* ¶ 79 ("the cardinal principles contained in the texts constituting the fabric of humanitarian law," the protection of civilians through distinction, and the prohibition of unnecessary suffering to combatants, are intransgressible principles of international customary law).

291 *Id.* ¶ 78 ("States do not have unlimited freedom of choice of means in the weapons they use").

"a use of force that is proportionate under the law of self-defense, must, to be lawful, also meet the requirements of the law applicable in armed conflict which comprise in particular the principles and rules of humanitarian law,"[292] thereby confirming that military necessity assessed under *jus contra bellum* cannot displace *jus in bello* constraints. That structural balance between military necessity and IHL is confirmed at the treaty level by the ICRC's updated Commentary on Common Article 3.

The ICRC's updated Commentary on Common Article 3 confirms this structural relationship at ¶ 561, establishing that where a provision of humanitarian law does not explicitly incorporate military necessity, "it must be presumed that the balance between military necessity and humanity has already been incorporated into the rule and thus military necessity may not be invoked to justify non-compliance with it," such that "military necessity arguments therefore do not justify acts or omissions inconsistent with the requirement of humane treatment"; subsequent ICTY jurisprudence reinforces this principle, the *Galić* Trial Chamber holding that Article 51(2) of Additional Protocol I "does not contemplate derogating from this rule by invoking military necessity" and rejecting as legally insufficient any framing of the civilian protection norm as one operative only "when not justified by military necessity.[293]

For autonomous cyber operations against critical infrastructure, therefore, environmental degradation (such as contamination of water systems or long–term damage to energy networks) and humanitarian consequences (including hospital failures and disruption of food

292 *Id.* ¶ 42 ("a use of force that is proportionate under the law of self-defense, must, in order to be lawful, also meet the requirements of the law applicable in armed conflict which comprise in particular the principles and rules of humanitarian law").

293 ICRC GC I Commentary, *supra* note 225, art. 3, ¶ 561 ("military necessity may not be invoked to justify non-compliance" with a humanitarian law rule that does not explicitly incorporate a military necessity qualification); Prosecutor v. Galić, Case No. IT-98-29-T, Trial Chamber I, Judgment, ¶ 44 (Int'l Crim. Trib. for the Former Yugoslavia Dec. 5, 2003) [hereinafter Galić Trial Judgment] (holding that AP I art. 51(2) "does not contemplate derogating from this rule by invoking military necessity"); *see also* Prosecutor v. Blaškić, Case No. IT-95-14-T, Trial Chamber I, Judgment, ¶ 180 (Int'l Crim. Trib. for the Former Yugoslavia Mar. 3, 2000) [hereinafter Blaškić Trial Judgment] (subsequently refined by Galić Trial Judgment ¶ 44 to eliminate any military necessity exception).

distribution) must be treated as integral components of proportionality and necessity analyses, not as external policy considerations weighed against military advantage.

Section II translates that normative imperative into the technical design requirements through which it becomes operationally enforceable.

II. Technical Safeguards and Resilient Architecture: Embedded Constraints on Autonomous Cyber Weapons

Existing law prohibits the effects Tier 3 systems structurally produce; it does not specify the design architecture through which compliance becomes achievable. Human Security by Design fills that gap. Operationalizing human security protection requires embedding technical safeguards directly into autonomous cyber weapons' design architecture and implementing resilient critical infrastructure capable of withstanding, detecting, and recovering from hostile autonomous cyber operations. The six safeguards analyzed in this Section are not engineering preferences; each is the direct translation of a specific IHL obligation into a design requirement. A system that lacks any one of the six cannot satisfy the treaty obligation from which that safeguard derives. A system that satisfies all six has been redesigned so as to no longer eliminate the human decision-maker IHL presupposes — it has crossed, in doctrinal terms, from the prohibited Tier 3 category into the conditionally permissible Tier 2 category established in Chapter 1. The following analysis takes each safeguard in turn, identifies the treaty obligation it operationalizes, and explains why the obligation cannot be discharged without it.

1. The Six Embedded Safeguards: Technical Architecture and IHL Grounding

Six core safeguards — kill switches, geofencing, time limits, target verification loops, comprehensive logging, and reversibility mechanisms — operationalize commanders' continuing obligations under AP I Articles 48, 51, 52, 57, and 58 to distinguish between military and civilian objects, ensure proportionality, and take all feasible precautions to minimize civilian harm.[294]

Kill switches provide remote deactivation capability, enabling human operators to terminate autonomous operations immediately upon detecting unintended effects, disproportionate civilian harm, or changing battlefield conditions invalidating original targeting assumptions, directly operationalizing the requirement under Article 57(2)(b) AP I that attacks "shall be cancelled or suspended" when causing excessive civilian harm.[295] That obligation is not static: Article 57(2)(b) AP I imposes a continuing duty to cancel or suspend attacks "if it becomes apparent" that they will cause excessive civilian harm imposes continuing obligations on commanders throughout attack execution, not merely at the planning stage; when autonomous weapons operate at extended ranges without direct human observation, "even greater caution is required" to detect violations and terminate operations. Kill switches operationalizing this duty require secure communication channels to prevent adversary interference; the U.S. Department of Defense mandates cybersecurity measures, including anti–tamper mechanisms and clear deactivation procedures for autonomous weapon systems.[296]

294 AP I, arts. 48, 51, 52, 57, 58; 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 1875, 2191, 2218; DoD Directive 3000.09, *supra* note 7, §§ 1.2(a)(1)(b), 1.2(a)(2)(c), 1.2(a)(3)(c).

295 AP I, art. 57(2)(b); 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 2220–2221; DoD Directive 3000.09, *supra* note 7, §§ 1.2(a)(1)(b), 1.2(a)(3)(c).

296 AP I, art. 57(2)(b); 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 2220–2221; DoD Directive 3000.09, *supra* note 7, §§ 1.2(a)(2)(a), 1.2(a)(3)(c).

The United States Department of Defense Nuclear Weapon System Surety Standards mandate positive measures to "prevent nuclear weapons involved in accidents or incidents, or jettisoned weapons, from producing a nuclear yield" and to "prevent inadvertent pre–arming, arming, launching, or releasing of nuclear weapons in all normal and credible abnormal environments," prioritizing safety over mission completion; this fail–safe principle is highly instructive for autonomous cyber weapons under Article 57's precautionary obligations requiring systems to terminate operations when they detect conditions causing disproportionate civilian harm.[297] Stuxnet provides a cautionary example: while the attack payload was specifically designed to sabotage Iranian nuclear centrifuges through manipulation of programmable logic controllers, the malware spread to approximately 100,000 infected hosts across over 155 countries, with Iran, Indonesia, India, and Azerbaijan constituting the most heavily affected States[298] demonstrating how autonomous cyber weapons can escape operational control and produce unintended transboundary effects that risk breaching IHL's distinction and proportionality requirements.[299]

Geofencing operationalizes Article 57's precautionary obligations by restricting autonomous weapons to pre–authorized operational domains through technical constraints. The U.S. Department of Defense mandates that autonomous weapon systems "complete engagements within a timeframe and geographic area, as well as other relevant

297 AP I, arts. 57(1), 57(2)(b); U.S. Dep't of Defense, *Instruction No. 3150.02, DoD Nuclear Weapon Systems Surety Program* §§ 1.2(c)(1), 1.2(c)(3), at 4 (Dec. 17, 2024) [hereinafter DoD Instruction 3150.02], https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/315002p.PDF (last visited Apr. 16, 2026).

298 Nicolas Falliere, Liam O Murchu & Eric Chien, *W32.Stuxnet Dossier Version 1.4*, at 6–7 (Symantec Corp. Feb. 2011), https://nsarchive2.gwu.edu/NSAEBB/NSAEBB424/docs/Cyber-044.pdf (last visited Apr. 16, 2026) [hereinafter Falliere, O Murchu & Chien, *Stuxnet Dossier*].

299 U.S. Dep't of Defense, *DoD Manual 3150.02, DoD Nuclear Weapon System Safety Program Manual* ¶ 3.b(1), (3) (Jan. 31, 2014) (Change 4, May 7, 2021) [hereinafter DoD Manual 3150.02], https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodm/315002m.pdf (last visited Apr. 16, 2026); AP I, arts. 57(2)(a)(ii), 57(2)(b); Falliere, O Murchu & Chien, *Stuxnet Dossier*, *supra* note 298, at 3, 5; Ralph Langner, *To Kill a Centrifuge: A Technical Analysis of What Stuxnet's Creators Tried to Achieve* 3–4 (Langner Grp. Nov. 2013), https://www.cs.yale.edu/homes/jf/Langner.pdf (last visited Apr. 16, 2026) [hereinafter Langner, *To Kill a Centrifuge*].

environmental and operational constraints, consistent with commander and operator intentions," requiring systems to "terminate the engagement or obtain additional operator input before continuing the engagement" when unable to comply with spatial boundaries. Implementation mechanisms include network boundary controls restricting operation to designated IP address ranges or network segments, digital authentication requiring cryptographic authorization for target engagement, and hardcoded geographic constraints preventing operations beyond authorized areas, technical safeguards ensuring compliance with IHL's requirements to "take all feasible precautions" in attack execution and to minimize incidental civilian harm.300 Geofencing contributes to implementing the requirement under Article 48 AP I that parties "shall at all times distinguish between the civilian population and combatants and between civilian objects and military objectives" and the mandate under Article 52(2) AP I that "attacks shall be limited strictly to military objectives" by restricting autonomous weapons to pre-designated operational areas within which the deployed commander has conducted a pre-operational IHL assessment confirming that identified objects within the zone satisfy the military objective criteria of Article 52(2) AP I, subject to the continuing presumption under Article 52(3) AP I of civilian status for objects whose dual-use character gives rise to doubt, thereby reducing the risk of engaging protected objects outside authorized military zones. Because military objectives are defined by their "nature, location, purpose or use" rather than geographic location alone, geofencing must be combined with target verification mechanisms to ensure full compliance with IHL's distinction requirements both inside and outside operational boundaries.[301] Geofencing addresses the particular challenge of cyber operations affecting interconnected civilian infrastructure by constraining autonomous weapons to specific network segments and

300 DoD Directive 3000.09, *supra* note 7, §§ 1.2(a)(1)(b), 4.c(2); AP I, art. 57(2)(a)(ii).

301 AP I, arts. 48, 52(2); 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 1863, 2011, 2018.

preventing lateral movement across segment boundaries, thereby implementing Article 57(2)(a) (ii)'s requirement to "take all feasible precautions in the choice of means and methods of attack with a view to avoiding, and in any event to minimizing, incidental loss of civilian life, injury to civilians and damage to civilian objects" and Article 51(5)(b)'s prohibition on attacks "which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated."

The operational content of "feasible precautions" under Article 57(2)(a)(ii) AP I, a standard whose textual indeterminacy has generated substantial scholarly debate, has been authoritatively elaborated in the ICRC's authoritative Commentary on Article 57 AP I as requiring assessment of those precautions "which are practicable or practically possible, taking into account all circumstances ruling at the time, including humanitarian and military considerations," a formulation that calibrates the precautionary obligation to operational context rather than imposing an absolute technological standard.[302]

Henderson provides the most granular analysis of what constitutes "feasible" precautions across different operational contexts, arguing that feasibility must be assessed not only by reference to the technical capability of the attacking force but also by reference to the 'overall circumstances ruling at the time[303] including intelligence availability, time pressure, and the nature of the target, a multi–factor approach that, when applied to autonomous cyber weapons, requires assessment of the weapons

302 AP I, art. 57(2)(a)(ii); Protocol on Prohibitions or Restrictions on the Use of Mines, Booby-Traps and Other Devices as Amended on 3 May 1996 art. 3(10), May 3, 1996, 2048 U.N.T.S. 93 [hereinafter Amended Mines Protocol II], https://www.un.org/en/genocideprevention/documents/atrocity-crimes/Doc.40_CCW%20P-II%20as%20amended.pdf (last visited Apr. 16, 2026.

303 AP I, art. 57(2)(a); Declaration of the United Kingdom upon Ratification of AP I ¶ (b) (Jan. 28, 1998), https://ihl-databases.icrc.org/en/ihl-treaties/api-1977/state-parties/gb (last visited Apr. 16, 2026).

system's own technical capacity to implement precautionary measures as a component of the feasibility analysis.[304]

Sassòli provides the most systematic doctrinal treatment of the feasibility standard governing the active precautionary obligations under Article 57, establishing that the duty to take all feasible precautions imposes a context–calibrated legal obligation that must be independently assessed in each concrete engagement rather than discharged through generalized pre–operational planning protocols. Those who plan or decide upon an attack must evaluate the circumstances prevailing at the actual moment of attack: the importance and urgency of destroying the target; the range, accuracy, and effects radius of available weapons; targeting accuracy conditions; the proximity of civilian objects; the possible release of hazardous substances; and the protection of the attacking party's own forces. Sassòli adds the 'availability and feasibility of alternatives' and 'the necessity to keep certain weapons available for future attacks on targets that are militarily more important or more risky for the civilian population."[305] Sassòli simultaneously acknowledges that these precautionary obligations, "more operational and precise than the proportionality rule," generate a structural verification problem that is internal to the IHL compliance framework: "the planning and decision–making process of commanders are by definition secret, and it is therefore difficult to know what they knew at the time of the attack or the alternatives they had available to them."[306] Applied to autonomous cyber weapons operating through machine–learning architectures, this epistemic limitation acquires systemic and legally decisive character: where the decision–making process is not merely confidential but entirely delegated to

304 Ian Henderson, The Contemporary Law of Targeting: Military Objectives, Proportionality and Precautions in Attack Under Additional Protocol I 161–65 (Martinus Nijhoff Publishers 2009) [hereinafter Henderson, *Law of Targeting*]; Alan Backstrom & Ian Henderson, *New Capabilities in Warfare: An Overview of Contemporary Technological Developments and the Associated Legal and Engineering Issues in Article 36 Weapons Reviews*, 94 Int'l Rev. Red Cross 483, 492–93 (2012), https://doi.org/10.1017/S1816383112000707 (last visited Apr. 16, 2026) [hereinafter Backstrom & Henderson, *New Capabilities*].

305 AP I, arts. 57(2)(a)(i)–(iii); Sassòli, *IHL*, *supra* note 251, MN 8.343.

306 Sassòli, *IHL*, *supra* note 251, MN 8.343.

pre–programmed algorithmic routines whose outputs cannot be interrupted or corrected by a supervising human agent during active operation, the *in concreto* feasibility assessment demanded by Article 57 AP I, calibrated to the circumstances prevailing at the actual moment of attack, not as modelled at the moment of programming, cannot be satisfied by anterior design alone, thereby rendering mandatory the continuous human oversight mechanisms proposed in this section.

The Tallinn Manual 2.0 elaborates these precautionary obligations in the cyber operations context with particular specificity: the obligation to do "everything feasible" to verify that objectives are lawful military targets may require gathering intelligence through network mapping or other processes in order to allow those responsible reasonably to determine the attack's likely effects, particularly on the civilian population or civilian objects; and where gathering sufficient information is not practicable or practically possible, the decision-maker may be required to refrain from conducting the attack altogether, or to limit its scope to only those components or capabilities of the target system with regard to which there is sufficient information to verify their status as lawful targets.[307] The Commentary to Rule 115 establishes at paragraph 5 that the "everything feasible" standard does not require "every conceivable action" but actions "practicable or practically possible taking into account all circumstances ruling at the time, including humanitarian and military considerations," and the Experts disagreed at paragraph 6 on whether the obligation requires active intelligence collection before each cyber operation or only requires refraining from attack where available information is insufficient, with the majority holding that the standard demands affirmative verification steps proportionate to the scale of anticipated effects and the availability of collection methods.[308] Rule 116 Commentary at paragraph 5 confirms unanimously that "the

307 AP I, art. 57(2)(a)(i); Tallinn Manual 2.0, *supra* note 18, r. 115 cmt. ¶¶ 4–5, at 478–79.

308 Tallinn Manual 2.0, *supra* note 18, r. 115 cmt. ¶¶ 4–5, at 478–79; *accord* AP I, art. 57(2)(a)(i).

interconnectivity of cyber infrastructure between military and civilian systems means that the obligation to minimize indirect as well as direct collateral damage affects not only the choice of the cyber means used, but also how it is employed."[309] Rule 119 Commentary at paragraph 7 addresses the cyber-specific challenge that "some cyber-attacks may be difficult to continuously monitor, thus making it practically difficult to know whether to cancel or suspend them," concluding that "this difficulty would heighten the degree of scrutiny that is merited during the planning and decision phases of the attack"[310], precisely the analytical basis for the ATO's ex ante interpretability component, which is designed to discharge at the planning phase the verification burden that Rule 119's monitoring difficulty prevents from being discharged at the execution phase. The Tallinn Manual 2.0's framework thus does not merely accommodate the ATO; it structurally requires it for cyber operations where real-time monitoring is architecturally foreclosed.

First, the principle of distinction applies to cyber–attacks: Rule 93, Commentary, paragraphs 1–3 confirm that the prohibition on directing attacks against civilian objects and the obligation to distinguish between civilian and military objects at all times extends to cyber operations qualifying as „attacks" within the meaning of Rule 92.[311]

Second, the proportionality rule prohibits a cyber-attack expected to cause excessive incidental civilian loss relative to the concrete and direct military advantage anticipated. Rule 113 Commentary, paragraphs 3–5 addresses how 'incidental damage' is assessed in cyber operations: the majority of Experts agreed that physical harm resulting from cyber operations qualifies as incidental harm, while significant debate arose over whether disruption of civilian cyber services short of physical damage, such as temporary loss of access to essential digital infrastructure,

309 AP I, art. 57(2)(a)(ii); Tallinn Manual 2.0, *supra* note 18, r. 116 cmt. ¶ 5, at 480.

310 AP I, art. 57(2)(b); Tallinn Manual 2.0, *supra* note 18, r. 119 cmt. ¶ 7, at 484.

311 Tallinn Manual 2.0, *supra* note 18, r. 93 cmt. ¶¶ 1–3.

constitutes 'damage to civilian objects' within the proportionality calculus.[312] The ICRC's 2026 position paper reinforces the case for extending proportionality analysis to the full range of reasonably foreseeable civilian effects, including indirect and cascading harms from cyber operations against interconnected infrastructure.[313]

Third, those who plan or decide upon a cyber–attack must take 'all feasible precautions in the choice of means or methods of warfare employed in such an attack, with a view to avoiding, and in any event to minimizing' incidental civilian harm: Rule 116, Commentary, paragraphs 1–5 specifies that 'feasible' in the cyber context requires assessment of the attacker's technical capability to limit collateral effects, including the availability of alternative cyber means that could achieve the same military advantage with lower risk of civilian harm, and that the selection of targets, timing of attack, and choice of specific cyber tools are all subject to the precautionary obligation; paragraph 5 further emphasizes that the interconnectivity of cyber infrastructure between military and civilian systems means that the obligation to minimize indirect as well as direct collateral damage affects not only the choice of the cyber means used, but also how it is employed.[314]

Time limits operationalize the mandate under Article 57(2)(b) AP I that "an attack shall be cancelled or suspended if it becomes apparent" that it would cause disproportionate civilian harm by requiring automatic deactivation after specified operational durations, forcing continuous reassessment of whether attacks remain proportionate under changed battlefield conditions; the U.S. Department of Defense's temporal constraint mandate, cited above in the geofencing analysis, applies with equal force to time limits, thereby ensuring compliance with the obligation to "take all feasible precautions in the choice of means and methods of

312 Tallinn Manual 2.0, *supra* note 18, r. 113 cmt. ¶¶ 3–5; ICRC Article 36 Guide, *supra* note 248, at 15–16.

313 ICRC, *AWS and IHL (2026)*, *supra* note 189, at 8 ("commanders and other users of AWS must . . . account for all reasonably foreseeable and relevant changes in circumstances between the AWS activation and the strike.

314 Tallinn Manual 2.0, *supra* note 18, r. 116 cmt. ¶¶ 1–5; AP I, art. 57(2)(a)(ii).

attack with a view to avoiding, and in any event to minimizing, incidental loss of civilian life" under Article 57(2)(a)(ii), as proportionality under Article 51(5)(b) depends on "the concrete and direct military advantage anticipated" at the time of attack rather than outdated assumptions.[315]

Target verification loops require human confirmation before autonomous cyber weapons engage objects protected under IHL or dual–use infrastructure serving essential civilian functions, operationalizing the presumption of civilian status under Article 52(3) AP I: "In case of doubt whether an object which is normally dedicated to civilian purposes, such as a place of worship, a house or other dwelling or a school, is being used to make an effective contribution to military action, it shall be presumed not to be so used.[316] Autonomous weapons with indefinite operational lifespans create unacceptable IHL risks because battlefield conditions change dynamically: adversary forces evacuate military installations replaced by civilian refugees; medical units establish field hospitals in buildings previously housing command centers; civilian authorities repurpose dual–use infrastructure exclusively for humanitarian relief. The requirement under Article 52(2) AP I that military advantage be assessed 'in the circumstances ruling at the time' imposes continuing verification obligations that autonomous systems cannot satisfy without periodic human reassessment and the authority to cancel or suspend operations under Article 57(2)(b).Technical implementation of the obligation under Article 57(2)(a)(i) AP I to "do everything feasible to verify" that targets are military objectives requires autonomous weapons to assess target classification reliability before engagement, with human

315 DoD Directive 3000.09, *supra* note 7, §§ 1.2(a)(1)(b), 1.2(a)(3)(c); AP I, arts. 51(5)(b), 57(2)(a)(ii), 57(2)(b); 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 2208–2209, 2220.

316 AP I, art. 52(3) (operationalizing the civilian-status presumption as a structural design requirement for autonomous cyber weapons: target verification loops must implement this presumption by requiring human confirmation before engagement of dual-use or protected objects, because automated targeting systems lacking human-in-the-loop verification capacity cannot satisfy the "in case of doubt" standard at the moment of target selection-the presumption is inoperative without a human decision-maker capable of applying it).

review mandated when systems cannot eliminate doubt about an object's civilian or military character.

The doctrinal framework for distinguishing civilian objects from military objectives in cyber–attack contexts must engage specifically with dual–use cyber infrastructure, systems serving both civilian and military functions simultaneously. Under Article 52(2) of AP I, once such an object satisfies the two–pronged test, effective contribution to military action and definite military advantage, it forfeits the protection accorded to civilian objects under Article 52(1), irrespective of its concurrent civilian utility.[317] Its concurrent civilian functions do not restore civilian protection but must be weighed as anticipated collateral damage within the proportionality calculus under Article 51(5)(b).[318] Where doubt persists as to whether an object normally dedicated to civilian purposes is making an effective contribution to military action, Article 52(3) requires that it be presumed civilian, imposing a verification obligation that autonomous targeting systems operating without real–time contextual assessment of network interdependencies cannot reliably discharge.[319]

While IHL does not prescribe specific numerical confidence thresholds, leaving technical standards to evolve with State practice and Article 36 weapons review processes, operational implementation must calibrate verification requirements to context–specific risks in accordance with the command under Article 57(1) AP I to take "constant care to spare the civilian population": higher certainty requirements for dual–use infrastructure whose destruction would severely impact civilian populations, highest when operating in proximity to objects under special protection (medical units, cultural property), reflecting the graduated precautionary framework inherent in IHL's distinction between attacks on

317 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 2018, 2023; Tallinn Manual 2.0, *supra* note 18, r. 100 cmt. ¶ 10.

318 AP I, art. 51(5)(b); Tallinn Manual 2.0, *supra* note 18, r. 101 cmt. ¶¶ 1, 3.

319 AP I, art. 52(3); ICRC, *AWS Selected Issues*, *supra* note 15, at 13–14.

unambiguous military objectives and attacks in circumstances where doubt exists triggering Article 52(3)'s presumption of civilian status.[320] When autonomous weapons cannot verify with sufficient confidence that potential targets are military objectives creating the "doubt" contemplated by Article 52(3), then Article 57(2)(a)(i)'s obligation to "do everything feasible to verify" target status requires human involvement in targeting decisions, and the cancellation duty under Article 57(2)(b) mandates continuing human authority to intervene. The ICRC Position on AWS operationalizes these treaty obligations by requiring "human–machine interaction" to ensure "effective human supervision, and timely intervention and deactivation" when autonomous systems encounter target classification uncertainty.[321] The Rome Statute's prohibition on "intentionally directing attacks against civilian objects" as a war crime under Article 8(2)(b)(ii) and the war crime of "intentionally launching an attack in the knowledge that such attack will cause... clearly excessive" civilian harm under Article 8(2)(b)(iv) establish criminal liability frameworks for autonomous weapons deployment where commanders possess the requisite mental element: under Article 30(2)(b), a person has intent where "in relation to a consequence, that person means to cause that consequence or is aware that it will occur in the ordinary course of events" thus, commanders who deploy autonomous weapons systems knowing (with awareness, not mere possibility) that the systems will attack civilian objects or produce clearly excessive civilian harm due to unreliable target classification satisfy the Statute's mental element. Article 28 command responsibility applies when commanders knew or "should have known" their forces would commit such crimes and failed to take necessary measures, such as implementing target verification loops to prevent violations, establishing individual criminal liability for deploying autonomous weapons lacking adequate verification

320 AP I, arts. 36, 52(3), 57(1), 57(2)(a)(i); ICRC Article 36 Guide, *supra* note 248, at 2; Davison, *AWS Under IHL*, *supra* note 177, at 14–16; DoD Directive 3000.09, *supra* note 7, § 4.1.c(3).

321 AP I, arts. 52(3), 57(2)(a)(i), 57(2)(b); ICRC Article 36 Guide, *supra* note 248, at 2.

safeguards when the commander possesses awareness that violations will occur in the ordinary course of the system's operation.[322] Target verification loops are therefore not merely good practice but a structural precondition for avoiding individual and command responsibility under Articles 8, 28, and 30.

Comprehensive logging of autonomous weapons targeting decisions is a structural prerequisite for operationalizing IHL's accountability architecture: Article 90(2)(c)(i) of AP I empowers the International Fact–Finding Commission to "enquire into any facts alleged to be a grave breach," while Article 53(1)(a) Rome Statute requires the ICC Prosecutor to consider whether available information provides a reasonable basis to believe that a crime within the jurisdiction of the Court has been or is being committed, both mechanisms dependent on available evidentiary records.[323] Article 87(1) imposes affirmative duties on commanders to "prevent and, where necessary, to suppress and to report to competent authorities breaches," presupposing records enabling commanders to monitor autonomous systems under their effective control, while Article 86(2) establishes superior responsibility where commanders 'had information which should have enabled them to conclude' that violations were occurring, creating legal incentives to maintain records as exculpatory evidence of proper supervision.[324] This convergence of investigatory, supervisory, and responsibility obligations establishes what Vincent Boulanin and Maaike Verbruggen term "traceability" the ability to trace the operation, performance, and effect of an autonomous weapons system back to people involved in its development and use, as a functional legal requirement: without such records, States cannot discharge Article 91's responsibility for "all acts committed by persons forming part of its armed forces," nor can the weapons review obligation under

322 Rome Statute, *supra* note 233, arts. 8(2)(b)(ii), 8(2)(b)(iv), 28(1)–(b), 30(2)(b); Elements of Crimes, *supra* note 413, arts. 8(2)(b)(ii), 8(2)(b)(iv), at 13–14; Lubanga Trial Judgment, *supra* note 322, ¶ 1011.

323 AP I, art. 90(2)(c)(i); Rome Statute, *supra* note 233, art. 53(1)(a).

324 AP I, arts. 86(2), 87(1).

Article 36 be operationalized for systems lacking traceability mechanisms.[325]

Reversibility mechanisms enable human operators to undo or mitigate autonomous cyber operations' effects, particularly where unintended civilian harm occurs, or battlefield conditions change, addressing the deploying State's continuing precautionary obligations under Article 57(2)(b) AP I, which requires that attacks be cancelled or suspended when they become apparent to be causing disproportionate civilian harm, and supplemented by the defending party's obligations under Article 58 to take precautionary measures protecting civilian populations under its control, individual civilians and civilian objects under their control against the dangers resulting from military operations," and complementary attacking party duties under customary IHL to minimize civilian harm through all feasible means.[326] Autonomous weapons systems incorporating irreversible effects, particularly when targeting dual–use infrastructure, must be assessed against the environmental protection obligations of AP I: Article 35(3) prohibits methods or means of warfare intended or expected to cause widespread, long–term, and severe damage to the natural environment, while Article 55(1) requires that care be taken to protect the natural environment against such damage, including a prohibition on methods expected to prejudice the health or survival of the population.[327] The ICJ has affirmed that "the environment is not an abstraction but represents the living space, the quality of life and the very health of human beings, including generations unborn";[328] that States "must take environmental considerations into account when assessing what is necessary and proportionate in the pursuit

325 AP I, arts. 36, 91; Vincent Boulanin & Marta Bo, *Three Lessons on the Regulation of Autonomous Weapons Systems to Ensure Accountability for Violations of IHL*, ICRC Humanitarian L. & Pol'y Blog (Mar. 2, 2023), https://blogs.icrc.org/law-and-policy/2023/03/02/three-lessons-autonomous-weapons-systems-ihl/ (last visited June 15, 2026) [hereinafter Boulanin & Bo, Three Lessons].

326 AP I, art. 58; ICRC CUSTOMARY IHL STUDY, *supra* note 18, r. 15.

327 AP I, arts. 35(3), 55(1).

328 Nuclear Weapons Advisory Opinion, *supra* note 39, ¶ 29.

of legitimate military objectives";[329] and that Article 35(3) and 55 'embody a general obligation to protect the natural environment against widespread, long–term and severe environmental damage,"[330] obligations that autonomous cyber operations producing long–term critical infrastructure degradation cannot satisfy without pre–operational assessment of cumulative and irreversible environmental effects.[331]

2. Resilient Infrastructure Design: Legal Obligations and Technical Standards

Article 58's defensive precautionary obligations require States to implement resilient infrastructure architecture, network segmentation, air–gapping, redundancy, and anomaly detection, capable of withstanding, detecting, and recovering from autonomous cyber operations.[332]

Network segmentation creates logical and physical boundaries separating civilian critical infrastructure from military networks, preventing cyber operations targeting military objectives from propagating into civilian systems through shared network connections. From an IHL perspective, segmentation operationalizes the requirement under Article 58(b) AP I that parties "to the maximum extent feasible... avoid locating military objectives within or near densely populated areas" and the requirement under Article 58(c) AP I to "take the other necessary precautions to protect the civilian population" while these provisions address physical proximity, their underlying principle of separating military and civilian assets to minimize civilian harm can be extended to cyber domains where "proximity" means network connectivity rather than geographic distance.[333]

329 *Id.* ¶ 30.

330 *Id.* ¶ 31.

331 Nuclear Weapons Advisory Opinion, *supra* note 39, ¶¶ 29–31.

332 AP I, art. 58(b)–(c); ICRC CUSTOMARY IHL STUDY, *supra* note 18, r. 22; Tallinn Manual 2.0, *supra* note 18, r. 121 cmt. ¶¶ 3, 12.

333 AP I, arts. 58(b), 58(c); Tallinn Manual 2.0, *supra* note 18, r. 121 cmt. ¶ 8.

Air–gapping ensures the complete physical separation of critical civilian infrastructure control systems from internet–connected networks through disabling all wireless capabilities, physically removing network interface cards, prohibiting removable media use, implementing supply chain security, and establishing strict physical access controls. The obligation under Article 58(c) AP I to 'take the other necessary precautions', permits States to consider the feasibility of protective measures, including operational costs and security tradeoffs while air–gapping may complicate routine maintenance and security updates, such costs may be justified where critical civilian infrastructure compromise would cause extensive civilian harm, consistent with the 'maximum extent feasible' standard under Article 58 and customary IHL Rule 22.[334] Redundancy ensures continuity of essential civilian services when primary infrastructure is compromised through geographically distributed backup systems immune from localized attacks, diverse technology stacks preventing common–mode failures, automated failover mechanisms detecting primary system compromise and seamlessly transitioning operations to backup systems, and sufficient excess capacity ensuring degraded operations continue delivering essential services during attacks. Article 58(c)'s obligation to 'take the other necessary precautions' extends to implementing redundant systems that prevent cascading infrastructure failures during cyber-attacks. Where critical civilian infrastructure, such as healthcare facilities, food distribution networks, or financial systems, depends on vulnerable single points of failure, redundancy can constitute a feasible precaution to protect civilians against the dangers posed by military cyber operations, consistent with the defensive precautionary framework under Article 58 and customary IHL Rule 22.[335]

334 AP I, art. 58(c); Jean-Marie Henckaerts & Louise Doswald-Beck, Customary International Humanitarian Law r. 22 (Cambridge Univ. Press 2005); Tallinn Manual 2.0, *supra* note 18, r. 121 cmt. ¶ 10.

335 AP I, art. 58(c); Jean-Marie Henckaerts & Louise Doswald-Beck, Customary International Humanitarian Law r. 22 (Cambridge Univ. Press 2005); *see also* Tallinn Manual 2.0, *supra* note 18, r. 121 cmt. ¶ 3.

Machine learning–based intrusion detection systems (ML–IDS) capable of recognizing autonomous weapon behavioral signatures, including rapid lateral movement, systematic network reconnaissance, algorithmic targeting patterns, or known exploitation techniques, operationalize the precautionary obligations under Article 58(c) AP I by enabling timely detection and response before significant civilian harm occurs, consistent with passive precautionary measures recognized under customary IHL Rule 22 and Tallinn Manual 2.0 Rule 121.[336] The Tallinn Manual Experts agreed that the protective rationale of Article 58 AP I, minimizing civilian exposure to the effects of attacks on military objectives, applies with equal force in the cyber domain where "proximity" manifests as network connectivity rather than geographic distance, extending to network monitoring, segmentation of civilian from military cyber infrastructure, and development of response capabilities preventing propagation of cyber-attacks into interconnected civilian systems, subject throughout to the "maximum extent feasible" qualification.[337]

Large-scale automated data analytics systems may simultaneously function as a detection infrastructure and a documentation record for IHL accountability purposes. The capacity of such systems to identify patterns, anomalies, and correlations "that would otherwise be invisible to the human eye" reinforces the technical dimension of Article 58's precautionary obligations; their evidentiary output directly enables the accountability mechanisms established under Article 90 AP I and Article 53(1)(a) of the Rome Statute.[338] Artificial intelligence-enabled capabilities have already demonstrated their potential to "help parties enhance (or evade) compliance with the law of armed conflict," and Big Data-

336 AP I, art. 58(c); Tallinn Manual 2.0, *supra* note 18, r. 121 cmt. ¶¶ 1–6, 11–13; Jean-Marie Henckaerts & Louise Doswald-Beck, Customary International Humanitarian Law r. 22 (Cambridge Univ. Press 2005.

337 Tallinn Manual 2.0, *supra* note 18, r. 121 cmt. ¶¶ 11–13; *see also* AP I, art. 58.

338 Beth Van Schaack, *Leveraging Big Data for LOAC Enforcement: Finding the Needle in a Stack of Needles*, *in* Big Data and Armed Conflict 291, 292 (Laura A. Dickinson & Edward W. Berg eds., Oxford Univ. Press 2023) [hereinafter Van Schaack, *Big Data*] ("discern non-obvious patterns, trends, anomalies, and correlations that would otherwise be invisible to the human eye"); AP I, art. 90(2)(c)(i); Rome Statute, *supra* note 233, art. 53(1)(a).

derived records may be "harnessed for lead and background purposes, to corroborate other evidence, to prove particular evidentiary points, and even to establish the elements of war crimes charges levied against a particular defendant."[339] The functional integration of detection and evidentiary preservation within a single technical architecture responds simultaneously to the *ex-ante* precautionary demand of Article 58 and the *ex-post* accountability demand of the Rome Statute, an alignment that Human Security by Design must deliberately operationalize.

III. Procedural Guarantees, Meaningful Human Control, and the Authorization Framework

The three-component ATO (as a matter of Article 31(1) VCLT systematic interpretation of Article 57(2)(a)(i) AP I, established in Chapter 3, Section II.7) advanced in Chapter 3 generates six procedural requirements that are developed, together with their normative foundations and institutional architecture, across the six sub-sections of this Section. Six procedural requirements flow from the ATO's three components. Section III.1 operationalizes the epistemic condition through the pre-deployment certification framework. Section III.2 establishes the normative foundations — the attack threshold, the Martens Clause, and the cognitive-agency standard — that govern all six requirements. Section III.3 operationalizes meaningful human control through the tiered authorization framework. Section III.4 addresses continuing supervision, mandatory documentation, and command responsibility. Section III.5 operationalizes the cancellation condition through kill-switch implementation requirements. Section III.6 applies the Bemba Gombo "necessary and reasonable measures" standard to procurement accountability.

339 Van Schaack, *supra* note 338, at 292 ("artificial intelligence-enabled capabilities can help parties enhance (or evade) compliance with the law of armed conflict"); *id.* at 314 ("Big Data can be harnessed for lead and background purposes, to corroborate other evidence, to prove particular evidentiary points . . . and even to establish the elements of war crimes charges levied against a particular defendant.

"Meaningful human control," for which no binding international definition currently exists, is understood in this chapter to require, consistently with the ICRC's position on autonomous weapon systems and the CCW GGE Guiding Principles, that human operators: (i) understand the context, conditions, and probable consequences of their decision to deploy a given autonomous system; (ii) exercise adequate supervision of the system's functioning throughout its operational period; and (iii) retain the technical and operational capacity to intervene, override, or deactivate the system at any point in its operational lifecycle.[340] This functional description operates as analytical shorthand for the convergent operational conditions under which AP I Article 52(3), 57(2)(a)(i), and 57(2)(b) can be satisfied.

1. Pre-Deployment Certification (Epistemic Condition): The Article 36 Review Obligation

Article 36 of AP I mandates that "In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party."[341] Article 36's application to autonomous cyber weapons follows four interpretive steps under Article 31(1)-(3) VCLT.

First, the ordinary meaning of "weapon, means or method of warfare" under Article 31(1) VCLT encompasses any instrument, technique, or operational method through which a party to an armed conflict seeks to injure, kill, or neutralize the adversary or to damage or destroy objects. Article 36 does not limit this to kinetic instruments: its text addresses the employment of "a new weapon, means or method," with the term

340 ICRC, *AWS and IHL (2026)*, *supra* note 189, at 1–2, 17–18; *Guiding Principles Affirmed by the GGE on Emerging Technologies in the Area of LAWS*, Annex III to Final Report of the 2019 Meeting of High Contracting Parties, Guiding Principle (b), U.N. Doc. CCW/MSP/2019/9 (Dec. 13, 2019) [hereinafter GGE Guiding Principles] ("human responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines").

341 AP I, art. 36.

"means or method" functioning as an independent, open–ended category that neither requires nor implies physical projectile delivery. An autonomous cyber system deployed to disable SCADA infrastructure controlling critical civilian systems, thereby producing physical effects on persons and objects, falls within the ordinary meaning of "means or method of warfare" on any faithful textual reading.[342]

Second, the contextual reading under Article 31(1) VCLT, which requires that a treaty provision be interpreted in light of its context - including the treaty's other articles read as a coherent whole - reinforces this scope: Article 35 AP I, in the same Part III Section on methods and means of warfare, prohibits methods and means of warfare of a nature to cause superfluous injury or unnecessary suffering and widespread, long-term and severe damage to the natural environment, categories that autonomous cyber operations causing irreversible infrastructure degradation can satisfy without involving kinetic delivery. Article 35's parallel text confirms that the drafters conceived 'means and methods of warfare' as a functional category defined by operational effect, not by physical delivery mechanism.[343]

Third, the object and purpose of Article 36 under Article 31(1) VCLT, read together with the contextual materials established under Article 31(2), is expressly preventive: the ICRC's Guide to the Legal Review of New Weapons confirms that Article 36's purpose is to prevent States from deploying weapons that would violate international law in all circumstances, by determining their lawfulness before they are developed, acquired or otherwise incorporated into a State's arsenal, rather than waiting for violations to occur and responding ex post. An interpretation that excluded cyber weapons from this preventive obligation because they were not contemplated at the 1977 drafting would defeat Article 36's object and purpose by creating an unregulated category of weapons

342 AP I, art. 36; Vienna Convention on the Law of Treaties, *supra* note 89, art. 31(1).

343 AP I, arts. 35(1)–(3); 1987 ICRC AP Commentary, *supra* note 18, ¶ 1475.

whose effects on civilians can exceed those of the conventional weapons Article 36 indisputably covers.[344]

Fourth, the supplementary means of interpretation under Article 32 VCLT, including the *travaux préparatoires* and the circumstances of the treaty's conclusion, confirm this scope. The CDDH records establish that Article 36 was drafted as a deliberately open-ended, technology-neutral obligation: the Rapporteur of Committee III defined the review requirement as directing States to "analyse whether the employment of a weapon for its normal or expected use would be prohibited under some or all circumstances," without limitation to weapons categories identifiable in 1977, and with the express understanding that "determination by any State that the employment of a weapon is prohibited or permitted is not binding internationally, but it is hoped that the obligation to make such determinations will ensure that means or methods of warfare will not be adopted without the issue of legality being explored with care." This purposive drafting history reinforces the textual and contextual reading: autonomous cyber weapons are precisely the category of technologically novel means of warfare that Article 36's drafters sought to bring within a preventive review obligation before operational deployment.345

This establishes, for High Contracting Parties to AP I, an affirmative, forward-looking treaty obligation to assess autonomous cyber weapons' structural capacity for IHL compliance before operational deployment. For States not party to AP I, the parallel substantive obligation derives from the customary prohibition on indiscriminate weapons confirmed as a norm of customary international law by the ICJ in the *Nuclear Weapons* Advisory Opinion. In either case, the Article 36 review obligation, as applied to autonomous cyber weapons through the ATO's three-

344 ICRC Article 36 Guide, *supra* note 248, at 3.

345 Diplomatic Conference on the Reaffirmation and Development of International Humanitarian Law Applicable in Armed Conflicts, *Official Records*, vol. XV, CDDH/215/Rev.1, at 269, ¶¶ 30–31 (Federal Political Department, Bern 1978), *cited in* 1987 ICRC AP Commentary, *supra* note 18, ¶ 1469 & n.10; *see also* ICRC Article 36 Guide, *supra* note 248, at 3–5.

component framework advanced in Chapter 2 as a matter of *de lege interpretanda*, requires States to identify and prohibit deployment of systems inherently incapable of satisfying the distinction, proportionality, and precautionary requirements of the applicable legal framework.[346]

The ICRC's 2006 *Guide to the Legal Review of New Weapons, Means and Methods of Warfare*, despite its non-binding character, has acquired the status of the authoritative practical elaboration of the obligation laid down in Article 36 of AP I, as reflected in the weapons review procedures established by several States.[347] The Guide operationalizes that obligation through a sequential four-part inquiry, requiring the reviewing authority to determine, first, whether the weapon is prohibited or restricted by applicable treaty or customary international law; second, whether it is 'of a nature to cause superfluous injury or unnecessary suffering' contrary to Article 35(2) AP I; third, whether it constitutes a means of warfare 'which cannot be directed at a specific military objective' or 'the effects of which cannot be limited as required by international humanitarian law,' contrary to Articles 51(4) and 51(5) AP I; and fourth, whether the weapon is expected to cause widespread, long-term, and severe damage to the natural environment contrary to Articles 35(3) and 55 AP I, a criterion directly engaged by autonomous cyber weapons whose cascade effects through critical infrastructure may produce long-term environmental degradation.[348] For autonomous cyber weapons, each of these criteria attaches to architectural features analyzed in Chapter 3: a system whose targeting logic is constitutively opaque cannot, in practice, be subjected to a meaningful assessment of its capacity to be

346 AP I, art. 36.

347 AP,; ICRC Article 36 Guide, *supra* note 248, at 2–3, 5–6 (confirming that the aim of AP I art. 36 "is to prevent the use of weapons that would violate international law in all circumstances . . . by determining their lawfulness before they are developed, acquired or otherwise incorporated into a State's arsenal," and documenting established national weapons review procedures in Australia, Belgium, the Netherlands, Norway, Sweden, the United States, France, and the United Kingdom).

348 ICRC Article 36 Guide, *supra* note 248, at 11–17 (§§ 1.2.1–1.2.2) (prescribing a sequential inquiry in which the reviewing authority determines first whether the weapon is specifically prohibited by treaty or customary international law (§ 1.2.1, at 12–14), and second whether it complies with the general IHL rules applicable to all weapons, including the prohibitions on weapons of a nature to cause superfluous injury or unnecessary suffering, on weapons incapable of being directed at a specific military objective, and on weapons whose effects cannot be limited as required by IHL (§§ 1.2.2.1–1.2.2.2, at 15–17)).

directed at specific military objectives, while a system whose propagation dynamics foreseeably generate cascade effects cannot credibly be assessed as capable of limiting its effects to lawful targets alone. The obligation under Article 36 thus operates, in the context of autonomous cyber weapons, as a pre–deployment filter: where the design of a system precludes assessment against these criteria, the review cannot be completed, and the weapon cannot lawfully be introduced into service. The treaty proposals in Chapter 6 are intended to give concrete effect to this pre-deployment screening function through mandatory disclosure duties and structured behavioral testing requirements.

The verification condition — that the discrimination logic of an autonomous cyber weapon must be auditable by a qualified third party — is operationalized at the institutional level through the IAWCO Technical Secretariat established in Chapter 6, Section IV.1, whose managed-access procedures resolve the tension between architectural disclosure necessary for compliance assessment and operational disclosure that national security considerations legitimately preclude

William Boothby's systematic analysis of the targeting cycle confirms this pre–deployment screening function and identifies the specific analytical steps that a legally adequate weapons review must encompass: assessment of the weapon's capacity to be directed at specific military objectives (distinction), assessment of the controllability of its effects (proportionality), and assessment of whether its design permits the taking of feasible precautions (precautionary obligation).[349] For autonomous cyber weapons, Boothby’s framework is particularly instructive in that it treats the weapons review not as a one-time administrative determination but as a continuing obligation requiring reassessment when the weapon is modified, when its operational environment changes, or when new information about its effects becomes available, a standard

349 AP I, art. 36; William Boothby, *Some Legal Challenges Posed by Remote Attack*, 94 Int'l Rev. Red Cross 579, 584–85 (2012) [hereinafter Boothby, *Remote Attack*].

that the algorithmic adaptability and environmental sensitivity of autonomous cyber systems makes practically impossible to satisfy without the embedded safeguards and logging mechanisms developed in Section II above.[350]

Article 36's customary status *vis–à–vis* non–AP I parties requires express acknowledgment. The United States, the world's principal developer of advanced autonomous weapons systems, is not a party to AP I and conducts weapons reviews as a matter of policy rather than treaty obligation; however, it is significant that DoDD 3000.09 (Jan. 25, 2023) expressly excludes autonomous and semi–autonomous cyberspace capabilities from its scope, confirming that no dedicated US policy framework currently governs autonomous cyber weapons reviews, while the Law of War Manual acknowledges that general law of war rules, including distinction, proportionality, and precaution, apply to all weapons with autonomous functions.[351] As a matter of progressive development of the law, the ICRC has argued that the obligation to assess the legality of new means and methods of warfare before deployment reflects a general principle of the law of armed conflict binding upon all States, a position grounded in the ICJ's affirmation in the Nuclear Weapons Advisory Opinion that the cardinal principles of IHL "are to be observed by all States whether or not they have ratified the conventions that contain them" because they constitute "intransgressible principles of international customary law."[352] This affirmation supports extending the review obligation by analogy to non–AP I parties, whose conduct remains governed by the parallel customary rules identified in the ICRC

350 Boothby, *Weapons and the Law of Armed Conflict*, *supra* note 350, at 347–49, 352–53; Damian Copeland, Rain Liivoja & Lauren Sanders, *The Utility of Weapons Reviews in Addressing Concerns Raised by Autonomous Weapon Systems*, 28 J. Conflict & Security L. 285, 298–99 (2023) [hereinafter Copeland, Liivoja & Sanders, *Weapons Reviews*].

351 DoD Directive 3000.09, *supra* note 7, §§ 1.1(b)(1), 1.2(a)(1), 1.2(c) (expressly providing that the directive "does not apply to: (1) Autonomous or semi-autonomous cyberspace capabilities"); U.S. Dep't of Defense, *Law of War Manual* §§ 6.5.9, 6.5.9.2, 6.5.9.3 (June 2015, updated July 2023) [hereinafter DoD, *Law of War Manual*], https://media.defense.gov/2023/Jul/31/2003271432/-1/-1/0/DOD-LAW-OF-WAR-MANUAL-JUNE-2015-UPDATED-JULY%202023.PDF (last visited Apr. 16, 2026).

352 AP I, art. 36; Nuclear weapons Advisory Opinion, *supra* note 39, ¶ 79.

Customary IHL Study.[353] Accordingly, as a matter of *de lege ferenda* grounded in the foregoing *lex lata* principles and their doctrinal development, this chapter applies the Article 36 review framework to all States developing autonomous cyber weapons, while maintaining analytical rigor by distinguishing between the treaty obligation (binding on the 174 High Contracting Parties to AP I) and the customary analogue (applicable to all States as a matter of general international law), and acknowledging that this extension to customary law remains contested by some States, including the United States, which has rejected the treaty obligation while voluntarily implementing its functional equivalent.[354]

This position must, however, be assessed in light of the most systematic examination of the evidentiary record currently available. *Natalia Jevglevskaja* has conducted a detailed survey of State practice and *opinio juris* and concludes that no evidence of 'extensive and virtually uniform' State practice supports the customary status of the Article 36 obligation in its full formulation, that is, the obligation to review new weapons at the earliest stage of the acquisition process as a matter of legal duty rather than policy discretion. The same analysis demonstrates that even a narrower obligation, to review weapons before fielding, is equally unsupported by the requisite quantum of State practice and *opinio juris*. This chapter accepts Jevglevskaja's conclusion regarding the current State of the evidence: the specific procedural obligation codified in Article 36, requiring review 'in the study, development, acquisition or adoption' of new weapons, has not, as of the present writing, attained the status of customary international law binding upon non–AP I

353 ICRC CUSTOMARY IHL STUDY, *supra* note 18, rr. 70–71.

354 *See* ICRC Article 36 Guide, *supra* note 248, at 5 (universal application argument); Nuclear Weapons Advisory Opinion, *supra* note 39, ¶ 79 ("intransgressible principles . . . to be observed by all States whether or not they have ratified the conventions that contain them"); DoD Directive 3000.09, *supra* note 7, § 1.1(b)(1) (U.S. voluntary functional equivalent, expressly excluding cyberspace capabilities); ICRC, *State Parties to the Additional Protocols*, https://ihl-databases.icrc.org/en/ihl-treaties/api-1977 (last visited Apr. 16, 2026) (174 High Contracting Parties to AP I as of 2026).

parties.[355] That concession does not, however, affect the central argument: the position advanced here does not depend on the customary status of Article 36 in its procedural formulation. The underlying substantive obligation, to ensure that weapons employed in armed conflict are capable of being used in conformity with the rules of IHL, is analytically distinguishable from the procedural mechanism by which compliance is verified. That substantive obligation derives not from Article 36 per se but from the logical structure of the substantive prohibitions on indiscriminate weapons under customary IHL Rules 11 and 12, Rule 11 prohibiting indiscriminate attacks, and Rule 12 defining such attacks as those not directed at a specific military objective, employing a method or means of combat that cannot be so directed, or employing a method or means whose effects cannot be limited as required by IHL, and from the ICJ's holding in the Nuclear Weapons Advisory Opinion that the 'cardinal principles' of IHL, including the prohibition on employing 'weapons that are incapable of distinguishing between civilian and military targets', 'are to be observed by all States whether or not they have ratified the conventions that contain them, because they constitute intransgressible principles of international customary law.[356]

A State bound by the customary prohibition on indiscriminate weapons is, as a matter of legal logic, required to determine, by whatever procedure it deems appropriate, that a weapon it introduces into service is not of a nature to violate that prohibition; to deploy a weapon without any prior assessment of its capacity to comply with these substantive constraints would be to accept knowingly the risk of employing an indiscriminate weapon, a risk that the substantive prohibition itself forecloses. Article 36 codifies one procedural modality for discharging this underlying substantive obligation; its non–customary status does not

355 Damian Copeland, *Book Review: Natalia Jevglevskaja, International Law and Weapons Review*, 13 J. Int'l Humanitarian Legal Stud. 379, 379–84 (2022), https://brill.com/view/journals/ihls/13/2/article-p379_009.xml (last visited Apr. 16, 2026).

356 Jean-Marie Henckaerts & Louise Doswald-Beck, Customary International Humanitarian Law rr. 11–12 (Cambridge Univ. Press 2005); Nuclear Weapons Advisory Opinion, *supra* note 39, ¶¶ 78–79.

relieve non–AP I parties of the substantive duty from which it derives. Accordingly, the *de lege ferenda* argument developed in this chapter, that all States developing autonomous cyber weapons should adopt structured pre–deployment review procedures, is grounded not in the customary status of Article 36 (which is absent) but in the customary status of the substantive rules whose effective implementation pre–deployment review serves.

The ICRC Commentary on AP I establishes at paragraph 1466 that the Article 36 review "is to be made on the basis of normal use of the weapon as anticipated at the time of evaluation," a standard the Rapporteur of Committee III elaborates at paragraph 1469: States must "analyze whether the employment of a weapon for its normal or expected use would be prohibited under some or all circumstances," and "a State is not required to foresee or analyze all possible misuses of a weapon, for almost any weapon can be misused in ways that would be prohibited," confirmed in the ICRC's own voice at paragraph 1480 as governing only "the normal use of the weapon as seen at the time of the evaluation, whether it were to be used in some or all circumstances (and not possibly misused)."[357] When transposed to autonomous cyber weapons, this review must assess four cumulative criteria. First, algorithmic discrimination capacity across all anticipated operational environments, including adversary electronic countermeasures and sensor deception. Second, proportionality reliability with respect to all reasonably foreseeable reverberating effects on civilian digital infrastructure, for which 'there are no temporal or geographic requirements… provided they are reasonably foreseeable.' Third, the availability of precautionary safeguards enabling 'timely intervention and deactivation, or, where necessary, self–destruction or self–neutralization,' subject to strict 'limits on the duration, scale and geographic scope of operation.' Fourth, sufficient system transparency to permit a lawful Article 36 determination at all,

357 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 1466, 1469, 1480.

since an AWS that precludes a human user from being able to understand, predict, or explain its output 'cannot be used in compliance with IHL,' rendering its use 'tantamount to blind firing.[358]

2. The Attack Threshold, the Martens Clause, and the Cognitive-Agency Standard for Meaningful Human Control

This section addresses three analytically distinct but sequentially ordered prerequisites for the authorization framework. Before the sub-sections that follow, the attack threshold under Article 49(1) AP I is confirmed as satisfied — as established in Chapter 3, Section I. Sub-section (a) identifies the Martens Clause as the residual normative baseline governing cases not covered by the specific treaty obligations examined in Sections III.1 and III.2 above. Sub-section (b) articulates the cognitive-agency standard.

The autonomous cyber operations analyzed in this Chapter satisfy the Article 49(1) AP I attack threshold — as established in Chapter 3, Section I — by virtue of their capacity to cause physical effects on infrastructure through functional harm to systems sustaining civilian life. The targeting rules of AP I engage in full.

a) The Martens Clause as Residual Normative Baseline

The Martens Clause, invoked as a contextual interpretive tool in the normative-tier analysis of Chapter 2, is here analyzed as a substantive normative baseline for cases not covered by the treaty provisions examined in this Section.

Prior to the specific *lex lata* obligations examined in this section, a residual normative baseline must be identified: the Martens Clause, codified in AP I Article 1(2) and the Preamble to Hague Convention (IV) Respecting the Laws and Customs of War on Land (1907), provides that "in cases not covered by this Protocol or by other international

358 ICRC Article 36 Guide, *supra* note 248, at 2, 9–10, 15–16.

agreements, civilians and combatants remain under the protection and authority of the principles of international law derived from established custom, from the principles of humanity and from the dictates of public conscience." The ICJ confirmed the Clause's continuing force in the *Nuclear Weapons* Advisory Opinion, holding that the Martens Clause, "reaffirmed in the first Article of AP I," is "merely the expression of the pre–existing customary law" binding all States, and that its "continuing existence and applicability is not to be doubted."[359] The ICRC's 2026 position paper, the most authoritative current institutional statement of how IHL applies to autonomous weapon systems, confirms that "existing IHL principles and rules, which are technology neutral, impose strict constraints on AWS and prohibit the development and use of certain types of AWS altogether," that it is essential to clearly articulate the types of AWS that cannot be employed in compliance with IHL, and specifically holds that "an AWS that precludes a human user from being able to understand, predict or explain its output cannot be used in compliance with IHL because its design, performance or operating features would render its use tantamount to blind firing.[360]

The express exclusion of autonomous cyberspace capabilities from the scope of DoDD 3000.09 constitutes, in doctrinal terms, an authoritative admission by the world's principal developer of such systems that no existing governance framework adequately addresses their specific IHL challenges, an admission that reinforces rather than undermines the Martens Clause analysis, for the governance gap thereby confirmed is precisely the "case not covered" by existing treaty or policy instruments within which the Clause's principles of humanity and dictates of public conscience supply the applicable normative standard. The command responsibility framework completes this chain: Article 28(1)(a)'s "knew

359 AP I, art. 1(2); Convention (IV) Respecting the Laws and Customs of War on Land pmbl., Oct. 18, 1907, 36 Stat. 2277 [hereinafter Hague Convention IV]; Nuclear Weapons Advisory Opinion, *supra* note 39, ¶¶ 84, 87.

360 ICRC Article 36 Guide, *supra* note 248, at 4, 9.

or, owing to the circumstances at the time, should have known" standard for military commanders, and Article 28(2)(a)'s 'consciously disregarded' standard for non-military superiors within the meaning of Article 28(2) of the Rome Statute, provide the individual accountability mechanisms through which this normative architecture generates enforceable legal obligations rather than aspirational principles.[361] These three normative layers, the Martens Clause's residual standard, Article 36's predeployment review filter, and the targeting rules' structural incompatibility with Tier 3 systems, form a coherent and mutually reinforcing architecture: where treaty specificity ends, the Martens Clause begins.

This conclusion must be distinguished from the counter–position advanced by Michael Schmitt, who argues that existing IHL provides sufficient constraints on autonomous weapon systems through obligations imposed on the persons deploying them rather than on the systems themselves, and that a finding of blanket incompatibility conflates IHL compliance with a technology–specific prohibition that existing law does not support.[362] The response to that position lies in the treaty text itself: Article 52(3), 57(2)(a)(i), and 51(5)(b) AP I do not merely impose obligations on deploying persons to be assessed retrospectively, they impose prospective conditions of lawfulness that must be satisfied before engagement, which is precisely what renders a system structurally incapable of meeting them unlawful per se rather than merely unlawfully deployed.[363] The ICRC's 2026 confirmation that autonomous weapon systems precluding human understanding, prediction, or explanation of outputs cannot be used in compliance with IHL expressly adopts the prospective conditions analysis and is irreconcilable with Schmitt's retrospective obligations model.[364]

361 Rome Statute, *supra* note 233, arts. 28(1), 28(2).

362 Michael N. Schmitt, *Fault Lines in the Law of Attack*, in Michael N. Schmitt, *Essays on Law and War at the Fault Lines* 175, 182, 188 (T.M.C. Asser Press 2012) [hereinafter Schmitt, *Fault Lines*].

363 AP I, arts. 51(5)(b), 52(3), 57(2)(a)(i).

364 ICRC Article 36 Guide, *supra* note 248, at 4–5, 9–10.

b) The Cognitive-Agency Standard for Meaningful Human Control

One doctrinal clarification must precede the tiered authorization analysis. The phrase "meaningful human control" functions in this chapter as analytical shorthand designating the convergent requirements of three specific treaty obligations: Article 57(2)(a)(i)'s obligation to do "everything feasible to verify" that objectives are neither civilians nor civilian objects before engagement; Article 52(3)'s requirement that doubt whether an object normally dedicated to civilian purposes is being used for military action triggers a presumption that it is not so used; and Article 57(2)(b)'s continuing duty to cancel or suspend attacks when it becomes apparent that the objective is not a military one or that civilian harm would be disproportionate.[365]

The precise doctrinal criterion that distinguishes legally sufficient human involvement in AI-assisted targeting from mere formal compliance has been elaborated by the Harvard Law School Program on International Law and Armed Conflict in its 2025 legal concept paper on cognitive agency. Lewis and Sweeney submit that under the existing international law applicable in armed conflict "it is assumed that, to administer the performance of IHL/LOAC obligations, humans need to exercise" what they term "cognitive agency," defined as "the undertaking and carrying out of a conscientious and intentional operation of mind by one or more humans vested with State legal capacity, through which that person or those persons implement the execution of the cognitive tasks demanded by the obligation."[366] The authors reason that this premise "arguably reflects a specification or an instantiation of existing

365 AP I, arts. 50(1), 52(3), 57(2)(a)(i), 57(2)(b).

366 Dustin A. Lewis & Hannah Sweeney, *Exercising Cognitive Agency: A Legal Framework Concerning Natural and Artificial Intelligence in Armed Conflict* 3 (HLS PILAC Legal Concept Paper, Jan. 2025), https://pilac.law.harvard.edu/exercising-cognitive-agency (last visited Apr. 16, 2026) [hereinafter Lewis & Sweeney, *Exercising Cognitive Agency*] ("By cognitive agency, we mean, with respect to administering the performance of an IHL/LOAC obligation, the undertaking and carrying out of a conscientious and intentional operation of mind by one or more humans vested with State legal capacity, through which that person or those persons implement the execution of the cognitive tasks demanded by the obligation.").

conditions of legality, not a new policy approach," grounding that assessment in analyses of how IHL/LOAC obligations are performed and how certain rules of State responsibility operate, a characterization that is analytically consistent with the *lex lata* framework rather than a *de lege ferenda* proposal.[367] Lewis and Sweeney further conclude that "the mere 'rubber stamping' by a human concerned" of an AI-related output would arguably not constitute an exercise of cognitive agency because it would lack an adequate conscientious and intentional operation of mind by that person." This conclusion maps directly onto the deficiency this Section identifies in Tier 3 autonomous cyber weapon targeting cycles: a Tier 3 operator who activates a system and observes its targeting outputs without independently executing the cognitive tasks demanded by Article 57(2)(a)(i), 52(3), and 57(2)(b) AP I exemplifies that disqualifying deficiency.[368] The functional description of "meaningful human control" employed in this chapter is accordingly coextensive with, and derives its operative content from, the cognitive-agency standard: any human involvement that fails to constitute a conscientious and intentional operation of mind implementing the cognitive tasks demanded by the relevant AP I obligations is legally insufficient, irrespective of nominal chain-of-command compliance or the formal presence of a human in the decision loop.[369]

367 *Id.* at 3 ("We reason that these premises arguably reflect a specification or an instantiation of existing conditions of legality, not a new policy approach. We ground that assessment in analyses of assumptions about executing cognitive tasks in connection with war, how IHL/LOAC obligations are performed, and how certain rules of State responsibility operate.").

368 Lewis & Sweeney, *Exercising Cognitive Agency*, *supra* note 366, at 19 ("In any event, under this second premise, the mere 'rubber stamping' by a human concerned of an AI-related output would arguably not constitute an exercise of cognitive agency because it would lack an adequate conscientious and intentional operation of mind by that person.").

369 *Cf.* AP I, arts. 52(3), 57(2)(a)(i), 57(2)(b); Lewis & Sweeney, *Exercising Cognitive Agency*, *supra* note 366, at 3.

3. Meaningful Human Control: Tiered Authorization, State Practice, and the Normative Basis of Tier 3 Prohibition

Legal requirements for meaningful human control necessitate tiered authorization frameworks calibrated to autonomy levels and target sensitivity. Tier 1 autonomy (human authorization of specific targets with automated execution) and Tier 2 autonomy (algorithmic target identification with mandatory human approval before engagement) can be considered legally permissible when accompanied by embedded safeguards specified in Section II, while Tier 3 autonomy (full autonomous operation without human supervision) is, as demonstrated in Chapter 3, structurally incompatible with existing *lex lata* obligations under Articles 52(3), 57(2)(a)(i), and 57(2)(b) AP I when targeting critical infrastructure, by reason of its structural incapacity to satisfy Article 52(3)'s presumption of civilian status for doubtful objects, Article 57(2)(a)(i)'s obligation to do everything feasible to verify that objectives are military objectives, and Article 57(2)(b)'s continuing duty to cancel or suspend attacks when disproportionate civilian harm becomes apparent.[370] The enhanced protection afforded to medical units under Article 12, cultural property under the 1954 Hague Convention, and works containing dangerous forces under Article 56 objects that "shall not be the object of attack," absent exceptional circumstances requiring highly contextualized legal determinations, necessitate individual human authorization before any engagement, as autonomous systems cannot reliably assess whether the narrow exceptions permitting attack apply in specific tactical circumstances.[371]

370 AP I, arts. 52(3), 57(2)(a), 58; ICRC Article 36 Guide, *supra* note 248, at 2, 7, 9.

371 AP I, arts. 12, 56; Hague Cultural Property Convention, *supra* note 275, art. 4.

4. Continuing Supervision, Mandatory Documentation, and the Command Responsibility Framework

Meaningful human control requires not merely ex ante authorization but continuous supervision throughout autonomous operations with technical capacity to override, modify, or terminate operations when circumstances change, deriving from Article 57(2)(b)'s requirement that "an attack shall be cancelled or suspended if it becomes apparent that the objective is not a military one or is subject to special protection or that the attack may be expected to cause" excessive civilian harm. This provision's "if it becomes apparent" language creates a continuing duty extending beyond initial targeting assessments: the obligation to cancel or suspend "applies not only to those planning or deciding upon attacks, but also and primarily, to those executing them," requiring commanders to monitor operations and intervene when new information emerges during the attack phase.[372]

To operationalize these continuing obligations, effective supervision requires: real–time telemetry transmitting autonomous weapon status, targeting decisions, and operational parameters to human supervisors; dashboard interfaces presenting complex operational data in formats enabling rapid human comprehension and decision–making under time pressure; override mechanisms allowing operators to modify targeting parameters, prohibit specific engagements, or terminate operations entirely; and graduated response options enabling proportionate interventions when partial corrections suffice.[373] After–action review protocols must assess autonomous system performance against IHL requirements, identifying systematic compliance failures requiring algorithmic retraining, operational restrictions, or weapons prohibition.

372 AP I, art. 57(2)(b); 1987 ICRC AP Commentary, *supra* note 18, ¶ 2220.

373 AP I, art. 57(2)(b); Boulanin et al., *Limits on Autonomy*, *supra* note 30, at 27–30.

Effective accountability architecture requires transparent documentation of development and deployment decisions, command responsibility extending to inadequate testing and supervision, and mandatory investigation of civilian harm incidents. While Article 36 itself does not mandate transparency, effective implementation of weapons review obligations requires documenting review processes, including testing methodologies, operational parameters, and algorithmic performance metrics, a position supported by the ICRC's guidance on legal review best practices. The comprehensive logging obligation established in Section II applies with equal force at the procedural supervision level: consistent with the traceability requirement, adequate documentary records are a prerequisite for triggering both the Article 53(1)(a) Rome Statute prosecutorial threshold and the individual criminal responsibility provisions of Articles 25 and 28 Rome Statute.[374]

Article 28 command responsibility attaches to the failure to take all necessary and reasonable measures within the commander's power to prevent the commission of crimes by forces under effective command and control. In the autonomous cyber weapons context, those measures encompass: requiring the completion of an Article 36 AP I weapons review before deployment authorization is granted; establishing technical safeguards, including kill-switches, target-verification loops, and geofencing, when feasible; implementing adequate testing protocols to assess discrimination reliability and proportionality compliance; and maintaining the capacity to cancel or suspend operations when monitoring reveals that civilian harm is apparent or disproportionate.[375]

374 AP I, arts. 86(2), 87(1), 90(2)(c)(i), 36, 91; Rome Statute, *supra* note 233, art. 53(1)(a); Boulanin & Bo, *Three Lessons*, *supra* note 325 (applying the traceability framework to establish comprehensive logging as a necessary incident of art. 36 weapons review: systems lacking mechanisms to record and trace autonomous targeting decisions create insurmountable barriers to the accountability architecture that AP I constructs-post-operation verification under art. 91, investigatory capacity under art. 90(2)(c)(i), and prosecutorial assessment under Rome Statute, *supra* note 233, art. 53(1)(a) all depend on available evidentiary records, rendering comprehensive logging a structural legal requirement that art. 36 weapons review procedures must verify before deployment authorization).
375 AP I, art. 36; ICRC, *AWS and IHL (2026)*, *supra* note 189, at 26–27; Rome Statute, *supra* note 233, arts. 28(1)–(b).

Article 28 Rome Statute establishes a bifurcated regime of superior responsibility. Under Article 28(1), a military commander bears criminal responsibility for crimes committed by forces under that commander's "effective command and control, or effective authority and control as the case may be," a standard requiring, in either form, the material ability to prevent or repress the commission of those crimes, which cannot be satisfied by mere influence over subordinates.[376] Under Article 28(2), a civilian superior incurs liability only where that superior 'consciously disregarded' information which clearly indicated" that subordinates were committing or about to commit crimes, a standard of conscious negligence or recklessness that is categorically distinct from, and set at a higher subjective threshold than, the constructive knowledge imposed on military commanders under Article 28(1).[377] This distinction carries direct significance for autonomous cyber weapons governance: military commanders directing operations involving autonomous systems are governed by the Article 28(1) "should have known" standard, while civilian officials responsible for authorizing and procuring such systems must be shown to have consciously disregarded information clearly indicating IHL incompatibility, a threshold that renders the absence of adequate a threshold that renders the absence of adequate Article 36 AP I legal review documentation legally significant evidence of that conscious disregard.[378]

376 Rome Statute, *supra* note 233, art. 28(1); Bemba Trial Judgment, *supra* note 233, ¶¶ 183–184.

377 Rome Statute, *supra* note 233, art. 28(2); Kai Ambos, Treatise on International Criminal Law, Vol. I: Foundations and General Part 227 (Oxford Univ. Press 2013) [hereinafter Ambos, *Treatise Vol. I*].

378 Rome Statute, *supra* note 233, arts. 28(1), 28(2); AP I, art. 36; Bemba Trial Judgment, *supra* note 233, ¶¶ 170(d), 196, 203.

5. Mandatory Embedded-Restraint Architecture (Cancellation Condition): Kill-Switch Implementation and Legal Exercise

The cancellation condition of the ATO, that kill-switch and override mechanisms must be technically implemented and legally exercised, generates the third procedural guarantee of this Section. This guarantee operates at the intersection of Article 57(2)(b)'s mandatory cancellation duty, and the technical safeguard architecture.[379]

Article 57(2)(b)'s obligation to cancel or suspend an attack "if it becomes apparent that the objective is not a military one" or that civilian harm would be excessive is not satisfied by the mere physical availability of a kill-switch mechanism. The obligation is legal in character: it requires that the human commander or operator possessing cancellation authority actually exercise that authority upon receiving information that triggers the cancellation duty. A kill-switch that is technically embedded but legally unexercised, because commanders lack the epistemic access to incoming targeting data necessary to recognize the trigger condition, fails the cancellation condition as a matter of treaty obligation, not technical design.

Three procedural requirements operationalize the cancellation condition at the institutional level. First, States deploying autonomous cyber weapons must establish, in writing, prior to deployment authorization, the specific trigger conditions under which the cancellation duty activates, defined by reference to operational parameters, civilian harm thresholds, and target-classification confidence levels, so that the commander responsible for cancellation can recognize those conditions in real time. Second, the chain of command must designate, by name and rank, the human officer bearing cancellation authority at each stage of the operational lifecycle, thereby preventing the diffusion of

379 AP I, art. 57(2)(b). The technical architecture for kill-switch implementation and override capacity is established in the preceding analysis of embedded safeguards.

responsibility that renders cancellation duties legally unenforceable in practice. Third, post-operation review must document whether cancellation triggers were encountered and, if so, whether the cancellation duty was exercised; failure to exercise a recognized cancellation trigger must be treated as a potential violation of Article 57(2)(b) AP I, subject to the mandatory investigation obligation under customary IHL Rule 158.[380]

As a matter of *de lege ferenda*, the Human Security by Design framework specifies the following technical implementation requirements for kill-switch and override mechanisms. These requirements are not derivable as direct, immediate obligations from AP I as currently formulated; they constitute proposed design standards whose adoption by States would satisfy the structural precautionary deficit identified in Chapter 2 and would reclassify a Tier 3 system as Tier 2 for the purposes of that Chapter's prohibition analysis.

The technical implementation requirements for kill-switches and override mechanisms are necessary but not sufficient to satisfy the cancellation condition; the procedural requirements in this subsection are sufficient. Together, they constitute the complete cancellation-condition compliance architecture that the ATO requires.[381]

Technical design constraints without procedural accountability mechanisms are aspirational; procedural accountability mechanisms without embedded technical constraints are unenforceable. Section IV addresses

380 AP I, art. 57(2)(b); Jean-Marie Henckaerts & Louise Doswald-Beck, Customary International Humanitarian Law r. 158, at 607–10 (Cambridge Univ. Press 2005) ("States must investigate war crimes allegedly committed by their nationals or armed forces, or on their territory, and, if appropriate, prosecute the suspects"; established as customary international law applicable in both international and non-international armed conflicts, grounded in the grave breaches regime of the Geneva Conventions, arts. 49, 50, 129, 146 respectively, and reaffirmed by the UN Security Council in relation to armed conflicts in Afghanistan, the Democratic Republic of the Congo, Kosovo, and Rwanda).

381 The distinction between necessary and sufficient conditions tracks the two-tier compliance architecture of this Chapter: the technical safeguards layer supplies necessary conditions; the authorization and accountability framework supplies sufficient conditions. Neither layer alone satisfies, as a matter of *de lege ferenda* design adequacy, the structural requirements that AP I arts. 52(3), 57(2)(a)(i), and 57(2)(b) generate-requirements which, under existing *lex lata* as established in Chapter 2, cannot be satisfied at all by Tier 3 systems and which can only be satisfied by Tier 2-reclassified systems meeting both layers of the Human Security by Design architecture.

the objections of pluralism, sovereignty, and military necessity that this integrated architecture must overcome to achieve universal adoption.

6. The "Necessary and Reasonable Measures" Standard and Procurement Accountability: The Bemba Gombo Framework

The contextual and circumstance-dependent nature of the 'necessary and reasonable measures' standard under Article 28 was authoritatively addressed by the ICC Appeals Chamber in *Prosecutor v. Jean-Pierre Bemba Gombo*, which reversed the Trial Chamber's conviction on the grounds that the standard is calibrated to the commander's actual means, information, and operational position, holding that an assessment of whether a commander took all necessary and reasonable measures 'must be based on considerations of what crimes the commander knew or "should have known" about and at what point in time,' and that Article 28 'only requires commanders to do what is necessary and reasonable under the circumstances.[382] The civilian procurement context is analytically distinguishable from the field command context in Bemba. Procurement officials operate under conditions of extended deliberation, with full access to Article 36 legal review files and without operational urgency, circumstances under which the information 'clearly indicating' IHL incompatibility required by Article 28(2)(b) Rome Statute is more likely to have been available to the official, making it correspondingly more difficult to deny conscious disregard where no adequate legal review was conducted.[383] These obligations acquire heightened practical significance in the specific context of the autonomous weapons deployment cycle, as the following concluding analysis confirms.

Civilian harm triggers mandatory investigation obligations under customary IHL, which should seek to determine: whether weapon

382 Bemba Appeals Judgment, *supra* note 233, ¶¶ 168–169 (¶ 168 establishing the knowledge-calibration principle; ¶ 169 establishing the "necessary and reasonable under the circumstances" formulation).

383 Rome Statute, *supra* note 233, arts. 28(1), 28(2); AP I, art. 36 (applying art. 28(2)'s conscious-disregard threshold to the civilian procurement context).

performance conformed to Article 36 legal review predictions; whether embedded safeguards functioned as designed; whether human supervision met Article 57 standards; whether civilian harm resulted from algorithmic errors, human judgment failures, or unforeseen cascading effects; and whether commanders or civilian superiors possessed the requisite mental element for criminal liability under Article 28(1)(a) or Article 28(2)(a) respectively.[384] The prospective governance architecture required by these accountability obligations finds its normative foundations not only in the Western IHL treaty tradition analyzed above but in legal frameworks whose temporal and relational conceptions of obligation supply resources that AP I's drafters did not develop.

IV. Normative Pluralism: A Methodological Note

The Human Security by Design framework rests exclusively on treaty and customary sources under Article 38(1) of the ICJ Statute. Five additional governance traditions — Haudenosaunee seven-generation thinking, Ubuntu's relational accountability doctrine, UNDRIP's participatory consent standard, Islamic Maslaha's precautionary principle, and buen vivir's lifecycle assessment framework — supply normative content for the *de lege ferenda* proposals of this chapter and Chapter 6. They are incorporated exclusively at the *de lege ferenda* plane, not as legally binding sources and not as interpretive tools under Article 31(3)(c) VCLT. Their inclusion reflects the judgment that a governance architecture regulating a technology deployed worldwide cannot, as a matter of political legitimacy and institutional durability, rest exclusively on Western legal positivism. None of these frameworks affects any *lex lata* finding in this dissertation.

384 AP I, art. 87(1); Jean-Marie Henckaerts & Louise Doswald-Beck, Customary International Humanitarian Law r. 158 (Cambridge Univ. Press 2005).

V. Implementation Architecture and Compliance Mechanism

The substantive and procedural requirements established in Sections II through IV generate their legal force only if accompanied by a credible institutional architecture capable of monitoring compliance and enforcing obligations.

Translating Human Security by Design's substantive requirements into operational reality requires robust implementation architecture encompassing monitoring, verification, enforcement, and compliance incentives, drawing on successful arms control precedents - the IAEA safeguards regime, Chemical Weapons Convention verification protocols, and UN Security Council sanctions practice - while adapting mechanisms to autonomous cyber weapons' unique technical and legal characteristics. Effective monitoring begins with transparency obligations requiring States to report on autonomous cyber weapons development, Article 36 review outcomes, and operational deployment, modelled on the IAEA's safeguards system under the NPT, which requires non-nuclear-weapon States to accept independent inspection rights to verify that nuclear material is not diverted to weapons purposes.[385]

Adapted to autonomous cyber weapons, an analogous monitoring body, the proposed IAWCO (developed in Chapter 6), operating under UN auspices, would receive State submissions in four categories. First, documentation of autonomous systems under development, including technical specifications (autonomy level per the three-tier typology, embedded safeguards, and target discrimination testing results). Second, Article 36 review files including legal determinations, operational constraints for Tier 2 systems, and prohibition decisions for Tier 3 systems. Third, deployment notifications specifying operational parameters

385 Treaty on the Non-Proliferation of Nuclear Weapons art. III(1), *opened for signature* July 1, 1968, 729 U.N.T.S. 161 (*entered into force* Mar. 5, 1970), https://treaties.un.org/doc/publication/unts/volume%20729/volume-729-i-10485-english.pdf (last visited Apr. 16, 2026) [hereinafter NPT]; Statute of the International Atomic Energy Agency arts. XII(A)(5)–(6), *opened for signature* Oct. 26, 1956, 276 U.N.T.S. 3 (*entered into force* July 29, 1957), https://www.iaea.org/about/statute (last visited Apr. 16, 2026) [hereinafter IAEA Statute].

(geographic scope, temporal limits, authorized targets). Fourth, incident reports detailing unintended civilian harm, algorithmic failures, and corrective measures implemented. Confidentiality protections modelled on IAEA practice would safeguard legitimately classified military information while ensuring sufficient transparency for compliance assessment.

Beyond routine monitoring, verification mechanisms must enable investigation of suspected violations through challenge inspections and technical expert review, with the Chemical Weapons Convention establishing the most intrusive verification regime in arms control history: Article IX permits any State Party to "request an on-site challenge inspection of any facility or location in the territory or in any other place under the jurisdiction or control of any other State Party" based on concern regarding possible non-compliance, employing "managed access" procedures balancing verification needs against States' legitimate security interests.[386] Challenge inspections for autonomous cyber weapons would require technical experts to examine: algorithmic architectures, assessing whether systems incorporate embedded safeguards claimed in Article 36 reviews; testing datasets and validation methodologies, verifying target discrimination performance meets claimed standards; operational logs from deployed systems, confirming compliance with authorization procedures and human supervision requirements; and incident investigation evidence, determining whether civilian harm resulted from algorithmic failures, inadequate safeguards, or command responsibility violations, with cybersecurity forensics specialists, autonomous systems engineers, and international law experts constituting interdisciplinary inspection teams possessing the competencies necessary for credible technical-legal assessments.

386 CWC, *supra* note 400, art. IX(8); *id.* Verification Annex pt. X, ¶ 41 (managed access procedures).

The IAEA managed-access mechanism resolves a verification challenge that is distinctive to software-based weapons: the tension between transparency sufficient to establish compliance and confidentiality sufficient to protect legitimate national security interests. The verification architecture's central analytical contribution is the distinction between architectural disclosure - which is necessary and sufficient for IHL compliance assessment - and operational disclosure - which is not required and may legitimately be withheld on national security grounds - a distinction that resolves the transparency-confidentiality tension identified above. Architectural disclosure, revealing the structure of an autonomous system's decision-making pipeline, the nature of its human authorization mechanisms, the design parameters of its propagation controls, and the results of standardized behavioral testing, is sufficient to enable the IAWCO Technical Secretariat to assess whether a declared system is compatible with the prohibition on Tier 3 operation and the weapons review criteria established in the proposed treaty's operative articles. Operational disclosure, revealing target lists, specific network objectives, classified intelligence assessments, or deployment timetables, is not required for this assessment and may legitimately be withheld under managed-access procedures. This distinction is the key that makes meaningful verification of software-based weapons systems feasible without compromising the legitimate security interests that States Parties must be able to protect if the treaty regime is to attract the participation of major cyber powers.[387]

This architecture would distribute institutional authority between the IAWCO Technical Secretariat's expert compliance assessments and the UN Security Council's enforcement competence under Chapter VII of

387 IAEA, *Guidance for States Implementing Comprehensive Safeguards Agreements and Additional Protocols*, IAEA Services Series No. 21, ¶¶ 6.3–6.5, 6.15–6.17 (2012) [hereinafter IAEA, *Guidance for States*], https://www.iaea.org/inis/collection/NCLCollectionStore/_Public/43/038/43038256.pdf (last visited Apr. 16, 2026); Trevor Findlay, *Confidentiality and Verification: The IAEA and OPCW*, Trust & Verify, No. 114, at 1–3 (May–June 2004) [hereinafter Findlay, *Confidentiality and VerificationConfidentiality and Verification*], https://www.vertic.org/media/assets/TV114.pdf (last visited Apr. 16, 2026).

the UN Charter, maintaining the separation between technical evaluation and political accountability that effective verification regimes require. The periodic review mechanism, modelled on the Montreal Protocol's requirement that Parties assess control measures beginning in 1990 and at least every four years thereafter in light of scientific, environmental, technical, and economic information, enables adaptive governance responsive to technological developments and operational experience while maintaining the institutional stability essential for a long-term compliance architecture.[388]

Critics may object that the proposed International Autonomous Weapons Control Organization represents politically unrealistic idealism given contemporary geopolitical realities, including great power competition between the United States, China, and Russia; Security Council paralysis demonstrated by deadlock over Syria's chemical weapons use and Russia's Ukraine invasion; and the absence of political will for new multilateral arms control institutions amid deteriorating international cooperation frameworks. This objection warrants serious engagement but proves ultimately unpersuasive for three related reasons. First, the IAWCO proposal explicitly acknowledges enforcement limitations inherent in the current international system, proposing graduated compliance mechanisms including regional organizations (European Union, African Union, Organization of American States) as complements to UN–centered enforcement, individual criminal accountability through International Criminal Court prosecutions when Security Council action proves impossible, and civil society monitoring through established non–governmental organizations possessing sophisticated verification capabilities independent of great power political constraints. Second, institutional proposals that initially faced dismissal as utopian have

388 Montreal Protocol on Substances that Deplete the Ozone Layer art. 6, *opened for signature* Sept. 16, 1987, 1522 U.N.T.S. 3 (*entered into force* Jan. 1, 1989), https://treaties.un.org/doc/publication/unts/volume%201522/volume-1522-i-26369-english.pdf (last visited Apr. 16, 2026) [hereinafter Montreal Protocol].

succeeded through sustained diplomatic engagement when normative foundations proved sufficiently compelling-the International Atomic Energy Agency encountered profound skepticism regarding verification feasibility during the 1950s yet became the cornerstone of nuclear non–proliferation architecture, while the OPCW's verification model, as noted above, achieved the most intrusive inspection regime in arms control history despite initial resistance from States protecting classified weapons programs. Third, international law scholarship performs an essential function in articulating normative frameworks that political actors may adopt when conditions permit implementation, as demonstrated by the Mine Ban Treaty and Cluster Munitions Convention precedents noted above, suggesting that Human Security by Design's institutional architecture may similarly provide the normative foundation for future regulatory action when autonomous cyber weapons proliferation creates sufficient political urgency for international cooperation.

The enforcement architecture, graduated sanctions, compliance incentives, and adaptive governance mechanisms through which these monitoring obligations acquire binding effect are developed in Chapter 6. Human Security by Design is the normative answer to the structural defect Chapters 2 and 3 identified. The IAWCO is the institutional mechanism that makes that answer enforceable. Where IHL's precautionary architecture failed because no human was present at the moment of lethal targeting, Human Security by Design relocates that human to the only moment where one still exists: the design phase, before the weapon is built, before it is deployed, before the law has already been violated.

Chapter 6
Governance Architecture and Institutional Frameworks for Autonomous Cyber Weapons

Human Security by Design, the normative framework of Chapter 5, is the red line's penultimate step. But a normative framework without an institution to verify and enforce it is aspiration, not law. Chapter 6 is what makes it law.

The argument this dissertation has constructed rests on one structural finding: IHL presupposes a human being at the moment of lethal targeting, and Tier 3 autonomous cyber weapons eliminate that human by design.

Chapters 2 and 3 established that this elimination defeats the precautionary architecture of AP I — in the *jus contra bellum* and in the *jus in bello* respectively.

Chapter 4 established that the same structural defect defeats the accountability frameworks of ARSIWA and the Rome Statute.

Chapter 5 responded with Human Security by Design — relocating IHL's obligations to the design phase, where a human still exists.

Chapter 6 builds the institution that makes that response enforceable. Without the IAWCO, Human Security by Design is a normative claim. With it, Human Security by Design is international law.

The structural analysis conducted across the preceding five chapters converges on a single institutional requirement: binding, verifiable, enforceable governance of autonomous cyber weapons. The existing governance landscape does not provide it. The CCW's engagement with lethal autonomous weapons systems, initiated through informal expert meetings in 2014 and institutionalized through the formal Group of Governmental Experts established by the Meeting of High Contracting

Parties in November 2016, has produced structured dialogue on LAWS but no binding instrument, its consensus requirement consistently preventing agreement beyond hortatory language. The OEWG process, established by GA Resolution 73/27, operates in parallel without coordination with the CCW track, generating duplicative State submissions and no binding output. Regional initiatives, the EU's 2023 autonomous weapons framework, the African Union's 2022 position paper, reflect genuine convergence on the prohibition of Tier 3 operation against civilians but lack an enforcement architecture. Bilateral arrangements between the United States and allied partners establish interoperability standards but not humanitarian constraints. The Secretary-General's 2024 report pursuant to GA Resolution 78/241 provides the most authoritative recent consolidation of State positions, confirming broad convergence on the need for binding regulation while deferring on the instrument's form. This chapter translates the normative foundation constructed in Chapter 5 into institutional architecture, advancing three propositions: that the existing governance landscape is constitutively insufficient; that a hybrid Security Council/General Assembly treaty approach provides the most legally coherent pathway to binding obligations; and that effective governance requires systematic integration of technology companies, civil society, and the research community into verification and enforcement architecture.

The argument proceeds in four stages, each building on the preceding chapters' doctrinal findings to produce institutionally precise and politically grounded governance proposals. Methodologically, Section I proceeds *de lege lata*, assessing the current governance architecture as a matter of existing positive international law; Sections II through IV advance proposals *de lege ferenda*, identifying the institutional architecture that the foregoing legal analysis requires as a matter of normative necessity. Section I assesses the existing governance landscape and establishes its constitutive insufficiency. Section II develops three treaty

options and defends the hybrid recommendation as the strategically optimal combination. Section III analyzes the indispensable governance roles of non–state actors. Section IV examines the specific challenges of verification and enforcement and proposes mechanisms calibrated to the distinctive opacity of software–based weapons systems.

I. Existing Governance Structures: Cyber Operations, Autonomous Weapons, and Institutional Deficits

The existing international governance architecture relevant to autonomous cyber weapons comprises three institutional tracks that have developed in partial isolation from one another: the United Nations GGE and OEWG processes addressing State behavior in cyberspace, the CCW GGE on lethal autonomous weapons systems, and regional, bilateral, and soft-law initiatives, including the Tallinn Manual process as a codification reference, the EU Cyber Diplomacy Toolbox, and ASEAN cyber norms frameworks.

1. The UN Group of Governmental Experts: Mandate, Achievements, and Limitations

The United Nations has pursued governance of State behavior in cyberspace primarily through two parallel expert processes - the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security (ICT GGE) and the CCW Group of Governmental Experts on Lethal Autonomous Weapons Systems examined in Section I.2 below. Neither body was designed to address autonomous systems and neither generates binding international obligations.

The Open–Ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security ('OEWG'), established by General Assembly Resolution 73/27 in 2018 to ensure universal participation beyond the GGE's restricted membership, produced a final report in 2021 that reaffirmed the

applicability of international law, emphasized confidence–building measures and capacity–building, and mandated a successor OEWG for 2021–2025.[389]

The OEWG's inclusive mandate, while normatively valuable in incorporating developing States' perspectives, has simultaneously constrained its analytical ambition: consensus requirements across all UN member States have produced instruments that affirm general principles without specifying the concrete obligations that effective governance of autonomous cyber weapons requires. The July 2025 OEWG final report reaffirmed existing frameworks but again declined to specify binding prohibitions applicable to autonomous systems.[390]

2. The CCW Process on Lethal Autonomous Weapons: Normative Stalemate and Prospective Reform

The CCW GGE on Lethal Autonomous Weapons Systems has affirmed since 2017 the continued applicability of IHL to all weapons systems and confirmed that human responsibility cannot be transferred to machines — a principle reaffirmed in the November 2024 Rolling Text.

These principles share the fundamental limitation of all CCW GGE outputs: non-binding, deliberately imprecise on the content of 'meaningful human control,' and structurally unable to address the cyber-domain-specific challenges this dissertation identifies.

389 G.A. Res. 73/27, ¶ 5, U.N. Doc. A/RES/73/27 (Dec. 5, 2018), https://undocs.org/A/RES/73/27 (last visited Apr. 16, 2026); U.N. Open-Ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security, *Final Substantive Report*, ¶¶ 8, 48, 51, 54–55, U.N. Doc. A/75/816 (Mar. 10, 2021), https://undocs.org/A/75/816 (last visited Apr. 16, 2026) [hereinafter OEWG, *Final Substantive Report*]; G.A. Res. 75/240, ¶ 1, U.N. Doc. A/RES/75/240 (Dec. 31, 2020), https://undocs.org/A/RES/75/240 (last visited Apr. 16, 2026).

390 U.N. Open-Ended Working Group on Security of and in the Use of Information and Communications Technologies 2021–2025, *Final Report*, ¶¶ 1, 3, 26, 32–33, 39, 40(b)(i), U.N. Doc. A/80/257 (July 11, 2025), https://undocs.org/A/80/257 (last visited Apr. 16, 2026) [hereinafter OEWG 2021–2025, *Final Report*].

3. The Tallinn Manual Process, Regional Initiatives, and the Limits of Soft–Law Codification

The Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations represents the most systematic scholarly effort to apply existing international law to cyber operations and constitutes the principal interpretive reference for State legal advisers and military practitioners in this domain.[391] The Manual's authority remains advisory: as a non-binding restatement of expert views produced without representation from China, Iran, or Russia, it cannot supply the lex specialis norms that autonomous cyber weapons governance requires.

The CCDCOE has announced a revision process to update the Tallinn Manual 2.0, with stated aims of incorporating State practice developments since 2017 and exploring 'new topics of importance to States' through a five-year revision process launched in 2021.[392] As of the date of submission of this dissertation, no revised text has been published. The normative gap this dissertation identifies accordingly remains unaddressed by any instrument of greater legal authority than the original Tallinn Manual 2.0 and cannot be closed by any revised edition of the Manual, whose mandate extends to the restatement of existing law, not its progressive development.

Regional initiatives supplement the global processes without substituting for them. The European Union’s Cyber Diplomacy Toolbox and 2020 Cybersecurity Strategy establish mechanisms for collective attribution and graduated diplomatic responses but contain no provisions

391 Tallinn Manual 2.0, *supra* note 18, at xxiii, xxv–xxvi; Niccolò Lanzoni, *The Role of Expert Groups in Shaping International Cyberlaw: A Case Study of the Tallinn Manuals and International Law-Making*, 30 J. Conflict & Sec. L. 341, 351–52 (2025) [hereinafter Lanzoni, *Expert Groups in Cyberlaw*], https://academic.oup.com/jcsl/article/30/3/341/8266494 (last visited Apr. 16, 2026).

392 NATO Cooperative Cyber Defence Centre of Excellence, *The Tallinn Manual*, https://ccdcoe.org/research/tallinn-manual/ (last visited Apr. 30, 2026) ("Emerging State practice and the taking of public positions on international cyber law by many States since the Manual's publication necessitates an update of the 2017 edition"); *see also* NATO Allied Command Transformation, *NATO Centres of Excellence, Cooperative Cyber Defense (CCD COE)* (Aug. 29, 2023), https://www.act.nato.int/article/nato-centres-of-excellence-cooperative-cyber-defense-ccd-coe/ (last visited Apr. 30, 2026).

specifically addressing autonomous weapons or the accountability gap.[393] ASEAN's cyber norms frameworks similarly address baseline peacetime conduct rather than autonomous weapons in armed conflict.[394] New instruments are accordingly a legal necessity.

II. Proposals for New International Legal Instruments: Three Regulatory Options

The international community's most recent formal expression of collective political will regarding lethal autonomous weapons systems is General Assembly Resolution 80/57, adopted on 1 December 2025 by the eightieth session of the General Assembly at its 52nd plenary meeting. The resolution preambularly affirms that "international law, including the Charter of the United Nations, international humanitarian law and international human rights law, applies in relation to autonomous weapons systems," and further reaffirms that "any weapon, including autonomous weapons systems, that cannot be used in compliance with international humanitarian law must not be used."[395] Its operative paragraph 3 calls upon the High Contracting Parties to the CCW "to work towards completing the set of elements for an instrument being developed within the mandate of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, with a view to future negotiations." The resolution further notes with appreciation the open informal consultations convened in 2025 pursuant to its

393 Council of the European Union, *Conclusions on a Framework for a Joint EU Diplomatic Response to Malicious Cyber Activities ("Cyber Diplomacy Toolbox")*, Council Doc. 9916/17, ¶¶ 4–5 (June 19, 2017), https://data.consilium.europa.eu/doc/document/ST-9916-2017-INIT/en/pdf (last visited Apr. 16, 2026) [hereinafter EU, *Cyber Diplomacy Toolbox*]; Council of the European Union, *Revised Implementing Guidelines of the Cyber Diplomacy Toolbox*, Council Doc. 10289/23, ¶¶ 16, 25–27 (June 8, 2023), https://data.consilium.europa.eu/doc/document/ST-10289-2023-INIT/en/pdf (last visited Apr. 16, 2026) [hereinafter EU, *Revised CDT Guidelines*]; European Commission & High Representative of the Union for Foreign Affairs and Security Policy, *The EU's Cybersecurity Strategy for the Digital Decade*, JOIN(2020) 18 final, at 15–18 (Dec. 16, 2020), https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=JOIN:2020:18:FIN (last visited Apr. 16, 2026) [hereinafter EU, *Cybersecurity Strategy 2020*].

394 ASEAN, *ASEAN Cybersecurity Cooperation Strategy 2021–2025* (Draft) ¶¶ 26–28, https://dig.watch/resource/asean-cybersecurity-cooperation-strategy-draft-2021-2025 (last visited Apr. 16, 2026) [hereinafter ASEAN, *Cybersecurity Strategy 2021–2025*]; *Chairman's Statement of the 32nd ASEAN Summit*, ¶ 6 (Apr. 28, 2018), https://cil.nus.edu.sg/wp-content/uploads/2019/02/2018-Chairmans-Statement-of-32nd-ASEAN-Summit.pdf (last visited Apr. 16, 2026).

395 G.A. Res. 80/57, *Lethal Autonomous Weapons Systems*, pmbl. ¶¶ 2, 7, U.N. Doc. A/RES/80/57 (adopted Dec. 1, 2025, distributed Dec. 5, 2025), https://undocs.org/en/A/RES/80/57 (last visited Apr. 16, 2026) [hereinafter G.A. Res. 80/57].

predecessor Resolution 79/62 of 2 December 2024 and acknowledges the CCW GGE rolling text as "a significant milestone in the work of the Group towards the fulfilment of its mandate."[396]

The structural deficiencies documented in Section I - institutional fragmentation, deliberate normative vagueness, and the categorical absence of binding obligations - establish that the governance gap cannot be closed through incremental reform of existing forums. Three distinct treaty options are available within the established architecture of international law–making, each with a different institutional vehicle, normative scope, and political feasibility profile, and each capable in principle of generating the binding, technically specific, and verifiable obligations that autonomous cyber weapons governance requires. They are assessed in turn against three criteria: normative coherence with the IHL framework established in Chapters 3 through 5; institutional adequacy to sustain the verification and enforcement architecture; and political realism assessed by reference to the demonstrated behavior of major cyber powers in analogous arms control contexts.

1. Option One: An AP to the Geneva Conventions

The most normatively integrated option is the negotiation of a new AP to the Geneva Conventions of 1949 addressing autonomous cyber weapons, operating as *lex specialis* within the existing humanitarian law framework. This approach has substantial precedential support: the 1980 Convention on Certain Conventional Weapons expressly provides for APs on new categories of weapons, and the 1977 Additional Protocols themselves demonstrate that States have adopted new treaty instruments to address the conduct of hostilities, including methods and means of warfare, in ways that the 1949 Geneva Conventions did not

396 G.A. Res. 80/57, *Lethal Autonomous Weapons Systems*, pmbl. ¶¶ 3, 11, operative ¶ 3, U.N. Doc. A/RES/80/57 (Dec. 1, 2025) (adopted 164–6–7, 52d plen. mtg., 80th Sess.), https://undocs.org/en/A/RES/80/57 (last visited Apr. 16, 2026); *see also* G.A. Res. 79/62, *Lethal Autonomous Weapons Systems*, operative ¶¶ 7, 9, U.N. Doc. A/RES/79/62 (Dec. 2, 2024), https://undocs.org/en/A/RES/79/62 (last visited Apr. 16, 2026).

fully operationalize. More directly, the CCW's Protocol IV on Blinding Laser Weapons provides a categorical-prohibition precedent of decisive structural significance. That instrument was adopted on 13 October 1995, before any State had employed blinding laser weapons in combat, notwithstanding confirmed development and testing programmes in the United States, China, and Russia, establishing that anticipatory prohibition of a weapons technology is both legally feasible and politically achievable before operational deployment crystallizes State practice that might otherwise complicate the normative consensus.[397] The Protocol's operative articles prohibit both the employment and the transfer of laser weapons "specifically designed, as their sole combat function or as one of their combat functions, to cause permanent blindness to unenhanced vision," while preserving the lawful use of laser systems whose blinding effect is merely incidental or collateral, a structural distinction between prohibited purpose-specific design and lawful incidental effect that the tiered regulatory architecture proposed in Chapter 5 directly replicates in the autonomous cyber weapons context.[398] The Protocol thereby demonstrates that anticipatory prohibition of inherently incompatible weapon technologies is achievable where the normative case is compellingly made, providing a direct institutional and doctrinal precedent for

397 Protocol on Blinding Laser Weapons (Protocol IV to the Convention on Certain Conventional Weapons), arts. 1, 3, Oct. 13, 1995, 35 I.L.M. 1218 [hereinafter Blinding Laser Protocol]; Int'l Comm. of the Red Cross, *Blinding Laser Weapons: Questions and Answers* (ICRC 1996) (noting confirmed development and testing by the United States, China, and Russia prior to adoption, with no documented operational deployment in armed conflict at the time of the Protocol's entry into force).

398 Blinding Laser Protocol arts. 1–3, *supra* note 397 (prohibiting employment and transfer of laser weapons "specifically designed . . . to cause permanent blindness to unenhanced vision" while expressly exempting "blinding as an incidental or collateral effect of the legitimate military employment of laser systems" under art. 3; the Conference adopted Protocol IV by consensus on October 13, 1995 before any State had employed such weapons in combat, establishing the anticipatory-prohibition precedent); *see* Louise Doswald-Beck, *New Protocol on Blinding Laser Weapons*, 312 Int'l Rev. Red Cross 272, 272–74 (1996).

a per se prohibition on Tier 3 autonomous cyber weapons deployed against objects protected under AP I Articles 12, 54(2), and 56(1).[399]

The substantive content of such a protocol must address four matters that existing treaty law leaves unresolved. First, it must establish legally precise definitions of "autonomous cyber weapon" and "meaningful human control," replacing the deliberately vague formulations of the CCW GGE process. The definition of autonomous cyber weapon should adopt a functional criterion encompassing two regulatory sub-categories, reflecting the taxonomy established in Chapter 1 and the differentiated legal framework. The first sub-category, a Tier 3 autonomous cyber weapon, is defined as a system that selects and engages targets through algorithmic processes without requiring human authorization of individual targeting decisions, and whose decision logic cannot be contemporaneously observed or overridden by a human operator during operation; systems satisfying this definition are subject to the treaty's prohibitory rule in respect of deployment against objects protected under AP I Articles 12, 54(2), and 56(1) and equivalent customary norms. The second sub-category, a Tier 2 autonomous cyber weapon, is defined as a system that executes targeting decisions algorithmically within pre-authorized parameters while retaining a human operator in a supervisory capacity who possesses the technical ability to monitor, interrupt, or override individual targeting outputs but whose intervention is not required for each individual engagement; systems satisfying this definition are not prohibited but are subject to the treaty's conditional authorization regime, which requires prior demonstration of embedded safeguard compliance, kill switches, geofencing, time limits, target verification loops,

399 Blinding Laser Protocol arts. 1, 3, *supra* note 397 (art. 1: prohibition on employment and transfer of laser weapons specifically designed to cause permanent blindness to unenhanced vision; art. 3: blinding as an incidental or collateral effect of the legitimate military employment of laser systems not covered by the prohibition); Milena Sterio, *Autonomous Weapons Systems and the Need to Update International Humanitarian Law?*, 57 Case W. Res. J. Int'l L. 301, 305 (2025) (identifying the Blinding Laser Protocol as establishing that pre-emptive weapons prohibition is politically achievable before operational deployment, and that the structural distinction between prohibited purpose-specific design and lawful incidental effect provides a regulatory template applicable to autonomous weapons systems).

comprehensive logging, and reversibility mechanisms, as specified in Chapter 5, Section II.1, as a prerequisite for lawful deployment, with compliance verified through the mandatory pre-deployment review procedure established in Article IV of the proposed IAWCO Convention, under which the Technical Secretariat conducts architectural disclosure review, standardized behavioral testing, and failure-mode assessment prior to issuing a deployment authorization certificate.[400]

The definition of meaningful human control should specify minimum technical and procedural requirements, pre–deployment verification of targeting parameters, continuous operational monitoring capacity, and verified deactivation mechanisms, rather than the abstract principle to which CCW deliberations have been confined. Second, the protocol should prohibit the deployment of Tier 3 autonomous cyber weapons against medical units (AP I, Articles 12–13), objects indispensable to the survival of the civilian population (AP I, Article 54(2)), and works or installations containing dangerous forces (AP I, Article 56(1)).[401] Third, mandatory Article 36 review requirements should be specified as procedural obligations with defined content: disclosure of machine–learning architecture, failure–mode assessment, and adversarial–robustness testing, with the resulting evaluations subjected to international peer review.[402] Fourth, compliance mechanisms, State reporting obligations, technical assistance for developing States, and interstate dispute

400 The designation of Article IV as the pre-deployment review provision follows the structural precedent of the Comprehensive Nuclear-Test-Ban Treaty art. II, *opened for signature* Sept. 24, 1996, 35 I.L.M. 1439 (not yet in force), Resolution of the States Signatories (Nov. 19, 1996) [hereinafter CTBT], https://www.ctbto.org/our-mission/the-treaty (last visited Apr. 16, 2026), art. IV(1), and the Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on Their Destruction art. III, *opened for signature* Jan. 13, 1993, 1974 U.N.T.S. 45 [hereinafter CWC], read together with the Verification Annex, Part II. The proposed IAWCO Article IV consolidates both functions-mandatory pre-deployment declaration and technical review-into a single article given the software-specific character of the regulated systems: unlike chemical agents susceptible to material sampling or nuclear tests detectable by seismic and radionuclide monitoring under the CTBT, art. IV(1)(a), the legally relevant characteristics of autonomous cyber weapons reside in AI models that may change through software updates and exhibit opaque behaviors in complex operational environments, making architectural disclosure review and standardized behavioral testing the only technically adequate verification modalities.

401 AP I, arts. 12(1), 13(1), 54(2), 56(1); CCW MSP, *2019 Report*, *supra* note 340, Annex III, Guiding Principles (b), (c); ICRC Article 36 Guide, *supra* note 248, at 7.

402 AP I, art. 36; ICRC, AI Military Domain Submission (2025), *supra* note 240, at 2-3, 9-10; Tallinn Manual 2.0, *supra* note 18, r. 110 cmt. ¶¶ 2, 6, 11–13.

resolution procedures should be established from the treaty's entry into force rather than deferred to subsequent protocols.

The principal limitation of this option is political. Amendment of the Geneva Conventions' framework through an AP requires adoption at a diplomatic conference and subsequent ratification by major military powers to achieve practical effectiveness; the United States, Russia, China, and key European Union member States are unlikely to accept binding prohibitions in the near term. This constraint is significant but not disqualifying. Protocol IV demonstrated that targeted prohibitions can be adopted and widely accepted even in relation to weapon systems that had not yet been used in combat, and that such prohibitions can contribute to the crystallization of a customary rule that binds all States, including non–parties, as part of the fundamental principles of international humanitarian law.[403]

2. Option Two: A Standalone Treaty on Autonomous Cyber Weapons

The second option is a comprehensive, standalone treaty specifically governing autonomous cyber weapons, modelled institutionally on the 1993 Chemical Weapons Convention and the regulatory architecture it established.[404] The CWC provides three institutional elements that an autonomous cyber weapons treaty should replicate and adapt.

First, it establishes the Organization for the Prohibition of Chemical Weapons with a Conference of States Parties as its principal organ charged with overseeing implementation and periodic review, an Executive Council responsible for promoting compliance between sessions, and a Technical Secretariat mandated to carry out verification measures.

---

403 Blinding Laser Protocol, *supra* note 397, arts. 1–3; Jean-Marie Henckaerts & Louise Doswald-Beck, Customary International Humanitarian Law r. 86, at 300–01 (Cambridge Univ. Press 2005); Nuclear Weapons Advisory Opinion, *supra* note 39, ¶ 79.

404 CWC, *supra* note 400, arts. I(1), VIII(1)–(2); Benjamin Perrin, *Lethal Autonomous Weapons Systems & International Law: Growing Momentum Towards a New International Treaty*, 29 ASIL Insights 1, 5–6 (Jan. 24, 2025) [hereinafter Perrin, *LAWS & International Law*], https://www.asil.org/insights/volume/29/issue/1 (last visited Apr. 16, 2026).

Second, its verification system combines mandatory State declarations of relevant programmes, routine inspections of declared facilities, and an intrusive challenge–inspection mechanism that any State Party may trigger to clarify suspected non–compliance and that must be accepted by the inspected State.

Third, its enforcement architecture empowers the Conference to adopt collective measures to ensure compliance, to restrict or suspend treaty rights in cases of persisting violations, and, in situations of particular gravity, to refer the matter to the United Nations Security Council.[405]

The technical challenge that differentiates autonomous cyber weapons from chemical weapons in the verification context must be confronted directly rather than minimized. Chemical weapons are physical objects amenable to material verification: inspectors can sample agents, measure stockpile volumes, and confirm destruction through observable processes. Autonomous cyber weapons are software–based systems whose legally relevant characteristics-particularly their targeting logic, decision parameters, and failure modes-reside in AI models that may change through software updates and exhibit opaque, unintuitive behaviors in complex operational environments. This distinction does not render verification impossible; it demands that verification methodology be adapted to the nature of the object being verified. Verification must prioritize access to algorithmic architecture documentation, rigorous standardized testing in conflict-relevant conditions, and iterative assessment using operational experience and proactive monitoring; the treaty instruments proposed in this Chapter operationalize that methodology through the IAWCO's Technical Secretariat.[406] Confidence–building measures,

405 CWC, *supra* note 400, arts. III(1)(a)(i)–(v), VIII(4), VIII(19), VIII(20), VIII(21)(k), VIII(31), IX(8), XII(1)–(4); Benjamin Perrin & Masoud Zamani, *The Future of Warfare: National Positions on the Governance of Lethal Autonomous Weapons Systems*, Lieber Inst. Article of War (Feb. 10, 2025) [hereinafter Perrin & Zamani, *Future of Warfare*], https://lieber.westpoint.edu/future-warfare-national-positions-governance-lethal-autonomous-weapons-systems/ (last visited Apr. 16, 2026).

406 ICRC, *AWS and IHL (2026)*, *supra* note 189, at 3–4; T.M.C. Asser Inst., *Iterative Assessment for Military Artificial Intelligence (AI) Systems* 5–6, 7–8, 9–10 (2025) [hereinafter T.M.C. Asser Inst., *Iterative Assessment*], https://www.asser.nl/media/so1pdprz/iterative_assessment_research_paper_2025.pdf (last visited Apr. 16, 2026).

including transparency on weapons–relevant programmes, systematic reporting of incidents and national implementation, and information–sharing arrangements calibrated to the specific risks of emerging weapons technologies, should supplement mandatory verification procedures.[407]

3. The Recommended Architecture: A Sequenced Hybrid of All Three Options

The third option, a binding resolution of the UN Security Council under Chapter VII of the UN Charter, offers the decisive advantage of immediate legal operability without requiring the extended negotiation that either a new protocol or a comprehensive treaty would entail. Chapter VII resolutions create binding obligations on all UN member States, and the Council has precedent for establishing regulatory frameworks in weapons–related contexts: Resolution 1540 of 2004 imposed binding non–proliferation obligations on member States in respect of weapons of mass destruction and their means of delivery, operating as what a number of scholars have characterized as a quasi-legislative instrument binding States without their individual consent.[408] A Chapter VII resolution establishing a moratorium on Tier 3 autonomous cyber-weapon deployments against protected civilian infrastructure, combined with a mandatory reporting obligation for development programmes, would impose immediate legal obligations without prejudicing the longer–term treaty negotiation process. The political feasibility of Option Three faces a structural obstacle: the same permanent member veto power that has foreclosed Chapter VII enforcement in analogous arms control contexts,

---

407 Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on Their Destruction art. V, *opened for signature* Apr. 10, 1972, 1015 U.N.T.S. 163 (*entered into force* Mar. 26, 1975) [hereinafter BWC]; *Working Paper on Providing Reassurance on BWC Implementation*, ¶¶ 4–5, U.N. Doc. BWC/CONF.VIII/PC/WP.34 (Aug. 16, 2016) [hereinafter BWC Working Paper]; Secretary-General, *Lethal Autonomous Weapons Systems: Report of the Secretary-General*, ¶¶ 28, 30, U.N. Doc. A/79/88 (June 30, 2024), https://docs.un.org/en/A/79/88 (last visited Apr. 16, 2026) [hereinafter UNSG, *LAWS Report 2024*].

408 S.C. Res. 1540, ¶¶ 1–2, U.N. Doc. S/RES/1540 (Apr. 28, 2004), https://digitallibrary.un.org/record/520326 (last visited Apr. 16, 2026); Paul C. Szasz, *The Security Council Starts Legislating*, 96 Am. J. Int'l L. 901, 901–02, 904 (2002) [hereinafter Szasz, *SC Legislating*].

discussed further in Section IV.3 below, would equally constrain the adoption of a binding Chapter VII resolution restricting Tier 3 autonomous cyber weapons among the permanent members themselves. A Chapter VII resolution imposing binding restrictions on Tier 3 autonomous cyber weapons is, therefore, as a matter of political analysis, unrealizable in the current geopolitical configuration. The value of Option Three lies not in its immediate feasibility but in its architectural utility as a fallback mechanism operable under rare conditions of great-power consensus, and as an institutional threat whose credibility may accelerate progress under the CCW track.

Even as a Chapter VI recommendation, the Security Council resolution would establish the normative precedent that autonomous cyber weapons targeting protected civilian objects are subject to international legal constraint — reducing the political cost of subsequent treaty ratification by removing the need to re-argue the principle. The CWC demonstrates that comprehensive arms control is achievable among geopolitical rivals when the normative case is sufficiently established.

The Security Council resolution and the General Assembly treaty process are the instruments through which that consensus is constructed and formalized. The treaty architecture is, however, structurally incomplete without the active participation of the actors who design, deploy, and monitor autonomous cyber weapons: Section III analyzes the indispensable governance roles of technology companies, civil society, and the research community whose engagement extends the regulatory framework beyond the State-centric plane to reach the engineering environments in which the legal challenge originates.

The integrated governance architecture this dissertation advances comprises three sequenced instruments.

First instrument (short term, zero to two years): A binding CCW Protocol establishing a per se prohibition on Tier 3 autonomous cyber weapon

deployment against objects protected under AP I Articles 12, 54(2), and 56(1), together with a mandatory Article 36 review obligation with defined procedural content. This provides the immediate normative floor that neither the General Assembly resolution track nor the standalone treaty process can deliver within the operational timeframe that the structural incompatibility identified in Chapter 2 makes urgent.

Second instrument (medium term, two to seven years): The IAWCO Convention, modelled on the CWC institutional architecture and incorporating the Technical Secretariat verification regime developed in Section IV. This serves as the primary normative instrument, replacing the CCW Protocol's categorical prohibition with the differentiated Tier 2/Tier 3 regulatory framework and extending governance to the full weapons lifecycle.

Third instrument (concurrent): A Security Council resolution under Chapter VII — or, where permanent member consensus is unavailable, a Chapter VI recommendation under Article 36(1) of the Charter — functioning as the normative catalyst that establishes the principle of international legal constraint before treaty negotiation is complete.

III. Non-State Actors in Autonomous Weapons Governance: Roles, Responsibility, and Accountability

Any treaty regime whose compliance architecture addresses only the obligations of States operates with a structural blind spot at precisely the point where regulatory intervention would be most effective: the commercial research and development environments in which autonomous targeting technologies originate, the civil society networks through which norm compliance is monitored, and the academic communities whose technical expertise is indispensable both to the design of verification standards and to the identification of emerging compliance failures. Three categories of non–State actor, technology companies, civil society organizations, and the research community, discharge governance functions that no State–to–State instrument can replicate, and whose institutionalization within the proposed IAWCO framework is accordingly a condition of the regime's effectiveness rather than an optional enhancement. The sections that follow analyze each category's specific governance role and translate the diffuse responsibilities imposed by existing soft–law frameworks into the concrete, enforceable obligations required by the three-stage governance architecture proposed in this dissertation: (a) in the short term, an Additional Protocol to the CCW prohibiting Tier 3 autonomous cyber weapons; (b) in the medium term, a standalone treaty establishing the International Autonomous Weapons Control Organization (IAWCO) as the primary verification body; (c) as a concurrent and supplementary measure, a UN Security Council resolution imposing interim obligations pending treaty entry into force.

1. Technology Companies: Corporate Due Diligence and Public–Private Governance

Chapter 1 established that the technologies underlying autonomous cyber weapons originate principally in commercial research and development environments: the machine learning classifiers, cloud computing infrastructure, and open–source neural network libraries that constitute the operational core of autonomous targeting systems are products of the technology sector, not the defense–industrial complex. Governance that addresses States while ignoring the engineering environments in which these systems are designed is governance with a structural gap at precisely the point where regulatory intervention would be most effective.

Technology companies bear a responsibility to respect human rights, including, under Guiding Principle 12 of the UN Guiding Principles on Business and Human Rights, the rights protected by international humanitarian law, throughout their operations and value chains, a normative expectation that does not, as a matter of *lex lata,* constitute a binding treaty obligation but which the proposed IAWCO framework would, upon its entry into force, translate into binding legal obligations subject to the enforcement architecture established in Section IV.[409]

Corporate actors operating within autonomous cyber weapons supply chains bear a dual responsibility under the U.N. Guiding Principles on Business and Human Rights, an internationally accepted framework that is not limited to peacetime operations and expressly governs corporate

409 Special Representative of the Secretary-General, *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework*, princs. 12–13, U.N. Doc. A/HRC/17/31 (Mar. 21, 2011) [hereinafter UNGPs], https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf (last visited Apr. 16, 2026); *endorsed by* Human Rights Council Res. 17/4, ¶ 1, U.N. Doc. A/HRC/RES/17/4 (June 16, 2011), https://digitallibrary.un.org/record/706796 (last visited Apr. 16, 2026).

conduct within, or linked to, armed conflict.[410] That framework, which the Human Rights Council endorsed unanimously on 16 June 2011, establishes, as a matter of soft law, what the Commentary to Guiding Principle 11 frames as an avoidance duty, to avoid infringing on the human rights of others through the enterprise's own activities, products, or services, and a remediation duty, to address adverse human rights impacts with which the enterprise is involved, characterized in the academic literature as negative and positive obligations respectively, obligations that encompass, at minimum, the avoidance of assisting States in committing human rights violations through data-enabled operations and the sharing of data capable of preventing humanitarian crises or war crimes.[411]

Human rights due diligence, requiring prior impact assessment, transparency on data collection policies, and compliance with international human rights standards in responding to State demands, constitutes the minimum standard of corporate conduct under Principle 17 of the U.N. Guiding Principles, applicable to all business enterprises wherever they operate, including companies in the communications and information technology sector that supply data, technology, or equipment enabling autonomous cyber operations.[412]

The legal status of the UNGPs requires clarification. They do not constitute binding treaty obligations; the Human Rights Council's endorsement on 16 June 2011 was a political rather than a legislative act. Their

410 Galit A. Sarfaty, *Corporate Data Responsibility*, *in* Big Data and Armed Conflict: Legal Issues Above and Below the Armed Conflict Threshold 276, 282–83 (Laura A. Dickinson & Edward W. Berg eds., Oxford Univ. Press 2023) [hereinafter Sarfaty, *Corporate Data Responsibility*]; UNGPs, *supra* note 409, princs. 11, 14 ("The responsibility to respect human rights is a global standard of expected conduct for all business enterprises wherever they operate"); Human Rights Council Res. 17/4, *supra* note 409 (endorsing the Guiding Principles unanimously).

411 UNGPs, *supra* note 409, princs. 11, 13 (requiring business enterprises to "avoid causing or contributing to adverse human rights impacts through their own activities"); Sarfaty, *Corporate Data Responsibility*, *supra* note 410, at 283.

412 Sarfaty, *Corporate Data Responsibility*, *supra* note 410, at 283–85; UNGPs, *supra* note 409, princ. 17 (requiring human rights due diligence comprising "a process to identify, prevent, mitigate and account for how enterprises address their adverse human rights impacts"); U.N. High Commissioner for Human Rights, *The Right to Privacy in the Digital Age* 15, U.N. Doc. A/HRC/27/37 (June 30, 2014) [hereinafter OHCHR, *Right to Privacy*] ("Reports also indicate that metadata derived from electronic surveillance have been analyzed to identify the location of targets for lethal drone strikes").

normative force operates through three channels. First, they reflect and consolidate the due diligence standards already applicable to corporations under domestic legal systems, standards that can engage civil liability in the courts of States where corporations are domiciled. Second, the ILC's ongoing work on crimes against humanity and corporate due diligence in armed conflict contexts suggests a trajectory toward binding corporate responsibility obligations in the medium term.[413] Third, and most immediately relevant to the present analysis, the UNGPs generate reputational and market pressure that functions as a quasi-regulatory mechanism in the technology sector, where investor ESG frameworks increasingly treat IHL compliance in defense contracting as a material risk factor, giving the due diligence obligation operational effect independent of its formally non-binding character.

Applied to autonomous weapons development, this responsibility generates a due diligence obligation for companies developing machine learning systems with foreseeable military applications: they are expected, as a matter of the UNGPs' due diligence standard and, de lege ferenda, under the proposed IAWCO treaty's Article 36-equivalent review obligations, to assess and mitigate the risk that their products will be deployed in ways that violate IHL. The scope of this obligation is calibrated by foreseeability and proximity. A company that develops general–purpose computer vision systems bears a materially lighter obligation than one that designs target-discrimination systems for deployment by armed forces. Where a company designs target-discrimination

413 Int'l Law Comm'n, *Draft Articles on Prevention and Punishment of Crimes Against Humanity*, art. 6(8), *in* Report of the International Law Commission on the Work of Its Seventy-First Session, U.N. Doc. A/74/10 (2019), *reprinted in* 2 Y.B. Int'l L. Comm'n, U.N. Doc. A/CN.4/SER.A/2019/Add.1 (Pt. 2) [hereinafter ILC Draft Articles on CAH] (art. 6(8): "Subject to the provisions of its national law, each State shall take measures, where appropriate, to establish the liability of legal persons for the offences referred to in this draft article. Subject to the legal principles of the State, such liability of legal persons may be criminal, civil or administrative"); *id.* art. 3(2) (obligations apply "whether or not committed in time of armed conflict," confirming applicability to the autonomous cyber weapons context irrespective of the existence of a formally declared conflict). The ILC Draft Articles on Prevention and Punishment of Crimes Against Humanity address State obligations and corporate liability for the enumerated acts under art. 2; the Draft Articles on the Responsibility of International Organizations (DARIO), *in* ILC, *Report of the International Law Commission on the Work of Its Sixty-Third Session*, U.N. Doc. A/66/10 (2011), address international organizations, not private corporations, and are not applicable in this context.

systems for deployment by armed forces with actual knowledge that the resulting system cannot satisfy IHL distinction requirements, that company's conduct may, depending on the applicable domestic legal framework, engage corporate civil or criminal liability under the applicable domestic law of the State of incorporation or the State of deployment for facilitating an internationally wrongful act, and may independently engage individual criminal liability under the Rome Statute's accessory liability standard in Article 25(3)(c), aiding, abetting, or otherwise assisting in the commission of a crime, where the individual actors within the company possess the requisite knowledge that their contribution facilitates the commission of the crime.[414]

The proposed treaty regime should translate these diffuse responsibilities into specific obligations: mandatory IHL impact assessments prior to defense contracts, disclosure of foreseeable weaponization risks to contracting States and to the proposed IAWCO Technical Secretariat, and public–private cooperation frameworks, modelled on existing cybersecurity information–sharing arrangements, that leverage the technical expertise companies possess but government verification bodies lack.[415]

The technology sector's voluntary commitments, while insufficient as a substitute for binding legal obligations, constitute a meaningful governance resource that treaty regimes should harness rather than dismiss. Industry codes of conduct on responsible AI development with military applications have demonstrated the capacity to alter internal corporate governance even without legal compulsion, creating a compliance

414 Rome Statute, *supra* note 233, art. 25(3)(c); Prosecutor v. Furundžija, Case No. IT-95-17/1-T, Trial Chamber II, Judgment, ¶ 249 (Int'l Crim. Trib. for the Former Yugoslavia Dec. 10, 1998) [hereinafter Furundžija], https://www.icty.org/x/cases/furundzija/tjug/en/fur-tj981210e.pdf (last visited Apr. 16, 2026).

415 AP I, art. 36; Laura Bruun, Marta Bo & Netta Goussac, *Compliance with International Humanitarian Law in the Development and Use of Autonomous Weapon Systems: What Does IHL Permit, Prohibit and Require?* 17–18 (SIPRI Mar. 2023) [hereinafter Bruun, Bo & Goussac, *IHL Compliance*], https://www.sipri.org/publications/2023/policy-reports/compliance-international-humanitarian-law-development-and-use-autonomous-weapon-systems-what-does (last visited Apr. 16, 2026).

infrastructure that a mandatory regime can build upon rather than construct from zero.[416]

2. Civil Society: Monitoring, Advocacy, and Prohibition Norm Entrepreneurship

Civil society organizations have historically served as norm entrepreneurs, in Martha Finnemore and Kathryn Sikkink's sense of actors who invoke principled ideas to persuade States of the need for behavioral change, in the autonomous weapons governance context, following the operational model that produced the Anti-Personnel Mine Ban Convention (1997) and the Convention on Cluster Munitions (2008).[417] The International Campaign to Ban Landmines, whose advocacy directly contributed to the 1997 Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on Their Destruction (Ottawa Treaty)'s negotiation on an accelerated timeline that bypassed traditional multilateral forums, provides the paradigmatic model of effective civil society engagement with humanitarian law development.[418] The Campaign to Stop Killer Robots has performed an analogous stigmatization function for autonomous weapons, successfully placing the issue on the international agenda and contributing substantially to the CCW GGE's conceptual work, while simultaneously

---

416 *REAIM 2023 Call to Action*, ¶¶ 5, 14, 16 (Feb. 16, 2023), https://www.government.nl/binaries/government/documenten/publications/2023/02/16/reaim-2023-call-to-action/REAIM+2023+Call+to+Action.pdf (last visited Apr. 16, 2026) [hereinafter REAIM 2023 Call to Action]; Alisha Anand & Harry Deng, *Towards Responsible AI in Defense: A Mapping and Comparative Analysis of AI Principles Adopted by States* 5, 7–8 (UNIDIR 2023), https://unidir.org/files/2023-02/Brief-ResponsibleAI-Final.pdf (last visited Apr. 16, 2026) [hereinafter Anand & Deng, *Responsible AI in Defense.*

417 Martha Finnemore & Kathryn Sikkink, *International Norm Dynamics and Political Change*, 52 Int'l Org. 887, 895–98 (1998) ("Norms do not appear out of thin air; they are actively built by agents having strong notions about appropriate or desirable behavior in their community"; defining the three-stage norm lifecycle of emergence, cascade, and internalization, and identifying persuasion as the characteristic mechanism of norm emergence); Human Rights Watch & Harvard Law Sch. Int'l Human Rights Clinic, *Losing Humanity: The Case Against Killer Robots* 1 (Nov. 2012) ("Human Rights Watch and IHRC believe that such revolutionary weapons would not be consistent with international humanitarian law and would increase the risk of death or injury to civilians during armed conflict. A preemptive prohibition on their development and use is needed.").

418 Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on Their Destruction, Sept. 18, 1997, 2056 U.N.T.S. 211 [hereinafter Ottawa Treaty], pmbl. & art. 1; Stuart Casey-Maslen, *Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on Their Destruction*, U.N. Audiovisual Library Int'l L., at 1, 3 (United Nations 2010) [hereinafter Casey-Maslen, *Ottawa Convention Commentary*], https://legal.un.org/avl/ha/cpusptam/cpusptam.html (last visited Apr. 16, 2026).

identifying, with considerable analytical precision, the structural limitations of that forum's non–binding output.[419]

A governance framework adequate to autonomous cyber weapons must institutionalize civil society participation rather than treat it as external pressure on intergovernmental deliberations. Three institutionalized roles are proposed. First, civil society organizations with relevant technical expertise should be formally accredited as observers in the IAWCO Conference of States Parties, with the capacity to submit technical analyses and to request investigatory procedures, an arrangement modelled on OPCW engagement with non–governmental technical experts.[420] Second, civil society's norm stigmatization function should receive formal institutional recognition: the treaty's preamble should explicitly acknowledge the role of civil society advocacy in the development of the norms it codifies, following the precedent of the Ottawa Convention and the Treaty on the Prohibition of Nuclear Weapons.[421] Third, independent AI auditing capacity-systematic technical assessment of autonomous weapons systems against defined IHL compliance benchmarks-should be developed through civil society and academic partnerships and formally recognized by the treaty regime as admissible evidence in compliance proceedings.

419 Campaign to Stop Killer Robots, *Annual Report 2023*, at 3, 6–7, 19 (Apr. 2023–Mar. 2024), https://www.stopkillerrobots.org/wp-content/uploads/2024/08/SKR-Annual-Report-2023_WEB_Final.pdf (last visited Apr. 16, 2026); Reaching Critical Will, *CCW Report*, vol. 11, no. 2, at 1, 6 (Mar. 14, 2023), https://reachingcriticalwill.org/disarmament-fora/ccw/2023/laws/ccwreport/16778-ccw-report-vol-11-no-2 (last visited Apr. 16, 2026).

420 CWC, *supra* note 400, arts. VIII(15), VIII(19), VIII(21)(h); OPCW, *Rules of Procedure of the Conference of the States Parties*, r. 33, Doc. C-I/3/Rev.2 (as amended) [hereinafter OPCW, *Rules of Procedure*], https://www.opcw.org/documents/2013/04/c-i3rev2/rules-procedure-conference-states-parties-organisation-prohibition (last visited Apr. 16, 2026).

421 Treaty on the Prohibition of Nuclear Weapons pmbl., *opened for signature* July 7, 2017, 3372 U.N.T.S. 1 (*entered into force* Jan. 22, 2021) [hereinafter TPNW], https://ihl-databases.icrc.org/en/ihl-treaties/tpnw-2017/preamble (last visited Apr. 16, 2026).

3. The Academic and Research Community: Norm Entrepreneurship and Epistemic Authority

The academic and research community constitutes, in Peter Haas's terminology, an epistemic community: a professional network sharing recognized expertise and causal beliefs whose capacity to shape State behavior through the provision of authoritative technical and normative analysis has been demonstrated, in the autonomous weapons governance context, through the Tallinn Manual process, the ICRC expert consultation series, and the UNIDIR research programme on lethal autonomous weapons systems.[422]

The research community occupies a position of unique governance significance: it simultaneously generates the technical capabilities that create the regulatory challenge and the expert knowledge necessary to address it. This dual character generates responsibilities that existing professional ethics frameworks inadequately address. The primary vehicle for embedding these responsibilities is the development and institutional adoption of ethical guidelines for AI research with foreseeable military applications, guidelines that go beyond general scientific integrity principles to address the specific risk that dual-use AI research contributes to the development of weapons systems incompatible with IHL.[423] Research universities, major AI laboratories, and scientific funding bodies should require researchers to conduct ex ante assessments of whether their work could enable violations of international humanitarian law, to disclose foreseeable military applications prior to publication, and to engage with IHL specialists before releasing research with significant autonomous weapons implications. The dual–use research of concern

422 Peter M. Haas, *Introduction: Epistemic Communities and International Policy Coordination*, 46 Int'l Org. 1, 3 (1992), at 3 (defining epistemic communities as "networks of professionals with recognized expertise and competence in a particular domain and an authoritative claim to policy-relevant knowledge within that domain or issue-area"; arguing that "control over knowledge and information is an important dimension of power).

423 Anand & Deng, *Responsible AI in Defense*, *supra* note 416, at 5–7; G7 Think7 Task Force, *G7-led Principles for Military AI Governance: Imperatives and Pathways*, at 4, 6–7 (2023), https://archive.think7.org/wp-content/uploads/2024/05/T7it_tf1_pb04.pdf (last visited Apr. 16, 2026) [hereinafter G7 Think7, *Military AI Principles*].

policies developed in the biosecurity field, which require institutional committee review of research that may reasonably be anticipated to generate capabilities of weapons relevance, provide an adaptable model.[424]

IV. Verification and Enforcement: Technical Challenges and Adaptive Mechanisms

Verification and enforcement are not supplementary refinements of a treaty regime whose normative content is established independently of them; they are constitutive of the legal obligation's practical force, as the Biological Weapons Convention's persistent compliance deficit - attributable directly to the absence of a dedicated verification mechanism - demonstrates with cautionary clarity. The verification methodology reorientation from physical detectability to functional assessability, introduced in the context of the standalone treaty option in Section II.2 above, is operationalized in this Section through four specific mechanisms adapted to software-based weapons systems, requiring a methodological reorientation from the detectability of weapons objects toward the functional assessability of weapons behaviors.

1. Technical Verification Challenges: Opacity, Attribution, and Adapted Mechanisms

Verification of compliance with autonomous cyber weapons obligations presents technical challenges that are qualitatively different from those encountered in any existing arms control regime and that require purpose–designed responses. The arms control verification canon, developed through the Nuclear Non–Proliferation Treaty, the CWC, and the Comprehensive Nuclear–Test–Ban Treaty, is premised on the physical detectability of regulated items: nuclear warheads leave isotopic signatures; chemical weapons stockpiles have measurable volume; weapons

424 AP I, art. 36; U.S. Office of Science and Technology Policy, *United States Government Policy for Oversight of Dual Use Research of Concern and Pathogens with Enhanced Pandemic Potential* §§ 1.1–1.3, 2.1–2.3(C), (F) (May 6, 2024) [hereinafter OSTP, *DURC Policy 2024*], https://bidenwhitehouse.archives.gov/wp-content/uploads/2024/05/USG-Policy-for-Oversight-of-DURC-and-PEPP.pdf (last visited Apr. 16, 2026].

tests produce seismic signatures. Software–based weapons systems possess none of these properties. A machine learning model trained for autonomous targeting is computationally indistinguishable from a commercial image classifier, occupies commodity hardware, can be transferred across international networks in seconds, and can be modified without a physical trace.

Four verification mechanisms, adapted to the distinctive properties of software–based systems, are proposed. First, mandatory declaration requirements should oblige States parties to disclose the existence, general architecture, and operational parameters of autonomous cyber weapons development programmes, with declarations subject to technical review by the IAWCO Secretariat. Confidentiality constraints need not defeat this mechanism: verification of whether a system's design is compatible with meaningful human control requirements can be achieved through architectural disclosure, revealing the structure of the decision–making pipeline and the nature of human authorization mechanisms, without disclosing target lists, operational parameters, or classified capabilities. Second, standardized behavioral testing protocols should require States to subject declared systems to IHL compliance testing in controlled simulation environments, with results reported to the Secretariat under confidentiality protections. The technical standards for such testing, distinction accuracy benchmarks, proportionality assessment consistency metrics, and adversarial robustness thresholds under distribution shift, should be developed by the Scientific and Technical Advisory Board in consultation with States parties and civil society technical experts.[425]

Third, confidence–building measures should supplement mandatory compliance procedures through transparency–enhancing practices that generate mutual assurance short of formal verification. Advance notification of autonomous cyber weapons testing, voluntary exchanges of

425 CWC, *supra* note 400, arts. III(1)(a)–(d), IV(3)–(5); Anand & Puscas, *LAWS Resource Paper*, *supra* note 266, at 11–13, 28–29.

information on national autonomous weapons policies, and regular technical briefings between States parties would materially reduce the misperception risk that systems operating at machine speed generate, a risk that Chapter 2 identified as a structural cause of escalation in cyber conflict contexts.[426] Analogous confidence–building measures under the 1986 Stockholm Document and subsequent OSCE instruments demonstrate that transparency frameworks can be established even among adversarial States when their shared interest in risk reduction is recognized.[427] Fourth, challenge inspections, enabling any State party to request a technical assessment of a suspected non–compliant system by the IAWCO Secretariat, should provide a mechanism for addressing credible allegations of violation that States are unwilling to raise through diplomatic channels. The CWC's challenge inspection procedure, though not yet invoked by any State Party, has been identified by compliance experts as an underutilized instrument whose activation remains essential to the credibility of the non–compliance response architecture; the same institutional logic applies to autonomous weapons.[428]

The structural durability of this verification architecture requires honest assessment against the three failure scenarios that the IAWCO's principal political vulnerability generates. First, where a P5 State refuses to submit mandatory declarations: the treaty regime cannot, in the absence of Security Council Chapter VII authorization, compel declaration compliance from a permanent member. The operative response is the

426 OSCE, *Vienna Document 2011 on Confidence- and Security-Building Measures* §§ 9, 16–17, 36–37, 40, FSC.DOC/1/11 (Nov. 30, 2011), https://www.osce.org/files/f/documents/a/4/86597.pdf (last visited Apr. 16, 2026) [hereinafter OSCE, *Vienna Document 2011*]; Matthew P. Shearer et al., *BWC Confidence-Building Measures: Increasing BWC Assurance Through Transparency and Information Sharing*, 43 Pol. & Life Sci. 1, 1–3 (2024), https://www.cambridge.org/core/services/aop-cambridge-core/content/view/4A4DB68CC05D706FF12C6C33CCF07252/S0730938424000091a.pdf (last visited Apr. 16, 2026) [hereinafter Shearer et al., *BWC Confidence-Building*].

427 *Document of the Stockholm Conference on Confidence- and Security-Building Measures and Disarmament in Europe*, ¶¶ 6–7, 29, 65–66 (Sept. 19, 1986), https://www.osce.org/sites/default/files/f/documents/5/d/41238.pdf (last visited Apr. 16, 2026) [hereinafter Stockholm Document 1986].

428 CWC, *supra* note 400, art. IX(8)–(9); Ralf Trapp, *Ensuring Compliance*, *in* Mina Rozei, Paul Walker & Daryl Kimball eds., *The Future of the Chemical Weapons Convention* 14, 18 (CWC Coal. & Arms Control Ass'n June 2025) [hereinafter Trapp, *Ensuring Compliance*], https://www.armscontrol.org/sites/default/files/files/Reports/CWCC-Report_2025_FINAL.pdf (last visited Apr. 16, 2026].

customary crystallization pathway - as the core prohibition norms accumulate ratification by middle-power States and consistent State practice, their binding force on non-declarants derives from customary international law rather than from the treaty, following the trajectory already observable with the Ottawa Convention's prohibition norms. Declaration refusal does not defeat the treaty; it excludes the refusing State from the verification and compliance-incentive architecture while accelerating the normative isolation dynamic that makes that exclusion politically costly. Second, where a P5 State invokes national security to block challenge inspection: the IAWCO Convention should, following the CWC model at Article VI(9), provide that national security objections may restrict access to specific technical details but not defeat the inspection in its entirety - the inspecting team retains the right to conduct managed-access review, with the inspected State responsible for demonstrating to the Technical Secretariat that the restricted information is not relevant to the compliance assessment. This mechanism cannot eliminate asymmetric information; it reduces it to a level the international community has accepted as the practical outer boundary of arms control verification, as the NPT context demonstrates. Third, where challenge inspection is structurally never activated because the political cost of triggering it is prohibitive: this failure mode, accurately identified in the CWC context by compliance experts including Zanders, does not render the mechanism worthless, its deterrence function operates independently of its activation rate, and the treaty's graduated enforcement architecture, under Tiers I through III, is designed to reach the compliance threshold that challenge inspection protects without requiring its activation in most cases.

2. Enforcement Mechanisms: Graduated Sanctions, Prosecution, and Compliance Incentives

The punitive dimension should operate through four graduated tiers, calibrated to the severity of the violation and designed to exhaust compliance-facilitation and administrative remedies before imposing economic measures or Security Council referral. The graduated sanctions architecture operates across four tiers.

Tier I (compliance facilitation) requires mandatory declaration of the non-compliant system by the State party; deployment of IAWCO Technical Secretariat technical assistance to support remediation; and a 90-day supervised remediation period before further enforcement action is initiated, reflecting the procedural model of the CWC Executive Council's graduated response practice under Article XII(1)–(4) CWC.[429]

Tier II (administrative sanction) entails suspension of the State's IAWCO voting rights for the duration of non-compliance; formal publication of a non-compliance finding in the IAWCO's annual report; and referral to the Technical Secretariat for enhanced monitoring, including mandatory reporting intervals and accelerated behavioral testing requirements.

Tier III (economic measure) authorizes targeted financial restrictions on the State's autonomous weapons development entities, modelled on the asset-freeze architecture of Security Council Resolution 1737 (2006), which established entity-specific rather than State-wide economic measures as the appropriate instrument in the weapons-control context,[430] to be recommended by the Conference of States Parties and implemented by member States without requiring Security Council

429 CWC, *supra* note 400, art. XII(1) (empowering the Conference to take compliance measures upon the recommendation of the Executive Council); *id.* art. XII(2) (authorizing restriction or suspension of a State Party's rights and privileges where it fails to fulfil a remediation request within a specified time); *id.* art. XII(3) (authorizing the recommendation of collective measures in conformity with international law where serious damage to the object and purpose of the Convention may result); *id.* art. XII(4) (requiring mandatory referral to the United Nations General Assembly and Security Council in cases of particular gravity).

430 S.C. Res. 1737, ¶ 12, U.N. Doc. S/RES/1737 (Dec. 23, 2006).

unanimity, a power exercised through the Conference's own constitutional authority under CWC Article XII, which authorizes collective measures "in conformity with international law" without conditioning such authority on Security Council approval. This architecture thereby preserves enforcement functionality in the face of permanent member veto power.[431]

Tier IV (Security Council referral) requires the Conference of States Parties to bring cases of material breach endangering the object and purpose of the Convention to the attention of the Security Council for Chapter VII action, following the mandatory referral procedure of CWC Article XII(4),[432] and to the States Parties in whose territory the alleged violations occurred, or whose nationals were involved, for consideration of referral to the International Criminal Court under Article 13(a) of the Rome Statute, or to the Security Council for referral under Article 13(b).[433]

The limitations of ICC prosecution in the autonomous weapons context must be acknowledged with the same intellectual honesty that Chapter 4 applied to the responsibility gap analysis. The structural accountability deficit, the absence of contemporaneous cognitive agency over specific wrongful acts, generates obstacles to prosecution that the doctrinal reorientation toward the deployment authorization stage significantly narrows but cannot eliminate. ICC enforcement is therefore a necessary but structurally limited element of the regime; its value resides principally in its deterrent function at the authorization decision point rather than in

431 CWC, *supra* note 400, art. XII(3) ("In cases where serious damage to the object and purpose of this Convention may result from activities prohibited under this Convention . . . the Conference may recommend collective measures to States Parties in conformity with international law.").

432 CWC, *supra* note 400, art. XII(4) ("The Conference shall, in cases of particular gravity, bring the issue, including relevant information and conclusions, to the attention of the United Nations General Assembly and the United Nations Security Council."). *Cf.* CWC, *supra* note 400, arts. XII(2)–(3).

433 Rome Statute, *supra* note 233, art. 13 (setting out the three exclusive triggers of ICC jurisdiction: State Party referral under art. 13(a) read with art. 14; Security Council referral under Chapter VII of the UN Charter under art. 13(b); and *proprio motu* Prosecutor investigation under art. 13(c) read with art. 15).

its retrospective accountability capacity.[434] In parallel, the ICJ's jurisdiction over inter–State claims provides a mechanism for States parties to seek an authoritative interpretation of IAWCO provisions and to resolve disputes concerning alleged non–compliance. Under the objective framework of international State responsibility, attribution of an internationally wrongful act to a State requires no proof of individual intent; it is sufficient that conduct is attributable to the State and constitutes a breach of an international obligation, distinct from, and analytically prior to, the strict–liability framework examined in Chapter 4, Section II.3.[435]

The proposed treaty regime should embed three categories of positive incentives. First, technical assistance and capacity–building programmes for developing States seeking to assess and regulate dual–use AI technologies, offered exclusively to States parties in compliance with their declaration obligations, create direct material inducements for treaty adherence and declaration compliance. Second, preferential access to non–military AI cooperation initiatives, coordinated through the IAWCO and associated scientific bodies, provides ongoing reputational and economic benefits tied to compliance status. Third, explicit recognition of leading States' governance contributions in treaty review conferences generates the reputational premium that reinforces the strategic interest in compliance maintenance. The combination of credible punitive measures and meaningful compliance incentives reflects the demonstrated architecture of the most successful arms control regimes, the CWC, the NPT, and the Ottawa Convention, which have achieved

434 Rome Statute, *supra* note 233, arts. 30(1), (3); Alessandra Spadaro, *A Weapon Is No Subordinate: Autonomous Weapon Systems and the Scope of Superior Responsibility*, 21 J. Int'l Crim. Just. 1119, 1124, 1127 (2023) [hereinafter Spadaro, *AWS and Superior Responsibility*], https://doi.org/10.1093/jicj/mqad025 (last visited Apr. 16, 2026).

435 Statute of the International Court of Justice arts. 34(1), 36(1)–(2), June 26, 1945, 33 U.N.T.S. 993 [hereinafter ICJ Statute]; ARSIWA, *supra* note 21, arts. 2, 4–5, 8, 11.

compliance rates that purely punitive enforcement would not have produced.[436]

3. The Veto Problem and Its Three Legal Responses

The IAWCO's central political vulnerability, the veto power of the three principal autonomous cyber weapons developers in the Security Council, has three partial legal responses, none of which resolve the political problem but each of which preserves the treaty's operational functionality in the face of it. The principal developers of autonomous cyber weapons, the United States, the Russian Federation, and the People's Republic of China, are permanent members of the Security Council and hold veto authority over Chapter VII enforcement measures. Three partial responses are available within the proposed legal architecture. First, the IAWCO Convention can embed a mandatory referral mechanism modelled on CWC Article VIII(36) and XII(4), which impose obligations on both the Executive Council and the Conference to bring cases of particular gravity directly to the General Assembly and Security Council, and activate the Uniting for Peace procedure of Resolution 377(V), paragraph 1, where Security Council action is blocked by veto, thereby preserving institutional functionality without requiring unanimity. Each of the three responses preserves the treaty's operational functionality without resolving the underlying political problem: the customary crystallization pathway, the middle-power treaty track, and the compliance-incentive architecture each operate independently of Security Council unanimity.

This chapter has translated the normative architecture developed in the preceding chapters into institutional proposals that are concrete, legally

436 CWC, *supra* note 400, arts. X(3), X(5), XI(2)(b), VIII(20); NPT, *supra* note 385, art. IV(2); Ottawa Treaty, *supra* note 418, arts. 6(4), 12(2)(d); Ralf Trapp & Cheng Tang, *Enhancing the Management and Enforcement of Compliance in the Regime Prohibiting Chemical Weapons* 11–12 (UNIDIR 2021) [hereinafter Trapp & Tang, *Compliance Regime*], https://doi.org/10.37559/WMD/21/CWC/01 (last visited Apr. 16, 2026).

precise, and calibrated to the political constraints within which international governance must operate. The existing governance landscape, fragmented across the GGE, OEWG, CCW GGE, and regional initiatives, and producing exclusively non–binding outputs, is structurally and by design incapable of addressing the legal challenges that autonomous cyber weapons present. Among the three treaty options examined, this dissertation recommends the temporally sequenced governance architecture set out below, which operationalizes the hybrid recommendation advanced in Section II.3 through sequential rather than simultaneous deployment of its constituent instruments. The first-order priority is the negotiation of a standalone treaty on autonomous cyber weapons (Option Two), negotiated through the General Assembly's Sixth Committee, with the IAWCO established by that treaty as its primary institutional expression. The Security Council resolution (Option Three) operates as a concurrent procedural precursor - its function is to establish the immediate normative precedent and mandate the preparatory commission, not to substitute for the standalone treaty's substantive obligations. Option One (a new AP to the Geneva Conventions) is not recommended as a primary or interim instrument for the reasons stated in Section II.1, though its normative content is compatible with and could be absorbed into the standalone treaty's operative articles. The CCW Protocol recommended in the Conclusion - distinct from Option One's AP to the Geneva Conventions - operates as a short-term transitional instrument during the five-year negotiation period. This sequencing - short-term CCW Protocol, concurrent Security Council resolution, medium-term standalone treaty as the primary instrument - is the only architecture whose legal authority exists now under Charter Article 13(1)(a) and Chapter VII without amendment of any existing international instrument. The framework's effectiveness depends further on the systematic integration of technology companies, civil society organizations, and the research community, whose participation extends the regulatory

architecture beyond the State–centric plane to reach the engineering environments in which the challenge originates. Verification mechanisms adapted to software's distinctive properties, and enforcement structures that combine graduated pressure with credible adjudicatory pathways, produce a coherent and self-reinforcing regulatory architecture. That architecture exists to serve one purpose: to ensure that the human being IHL presupposes at the moment of lethal targeting is never permanently absent — that before a Tier 3 weapon is built, before it is deployed, and before the law has already been violated, a human decision-maker capable of bearing legal responsibility has certified that the system is not beyond human control. The IAWCO is the institution that makes that certification possible. Human Security by Design is the framework that makes it necessary. Together, they are the answer that the structural defect identified in Chapters 2 and 3 demands.

Conclusion

International humanitarian law cannot govern Tier 3 autonomous cyber weapons. Not because its rules are wrong. Because its rules presuppose something those weapons eliminate by design: a human being, present at the moment of lethal targeting, capable of verifying a target, weighing proportionality, and cancelling an attack. When that human is gone, the law has no one to speak to. The compliance architecture that IHL constructs over centuries of humanitarian development depends, at every operative moment, on the presence of that human. Tier 3 autonomous cyber weapons make that presence architecturally impossible.

That is this dissertation's finding. It is not a criticism of IHL. It is a diagnosis of what IHL cannot reach.

The Introduction posed the question in the form of a hypothesis: that international humanitarian law, as currently constructed, is structurally incompatible with Tier 3 autonomous cyber weapons — not because its rules are wrong, but because its compliance architecture presupposes a human being present at the moment of lethal targeting, and Tier 3 autonomous cyber weapons eliminate that human by design.

The six chapters that precede this Conclusion have confirmed that hypothesis. The Conclusion states what that confirmation means.

I. The Hypothesis Confirmed: Three Propositions Established

The hypothesis rests on three interlocking propositions. Each has been established.

The first proposition concerned the precautionary architecture of Additional Protocol I. The verification obligation of Article 57(2)(a)(i), the cancellation obligation of Article 57(2)(b), and the weapons-review obligation of Article 36 are not merely inadequately applied when confronted with Tier 3 autonomous cyber weapons. They are structurally incapable of operational implementation. The four technological characteristics established in Chapter 1 — propagation dynamics, temporal persistence, technical irreversibility, and attribution opacity — jointly generate a condition this dissertation designates verification impossibility: the state of affairs in which no human decision-maker, however diligent, can verify at the operative moment of force whether the object being targeted satisfies Article 52(2)'s military objective criteria, whether the expected incidental harm satisfies Article 51(5)(b)'s proportionality requirement, or whether the cancellation trigger under Article 57(2)(b) has been met. This is not a temporary gap that better technology will close. It is an architectural consequence of the deep-learning paradigm whose effectiveness at pattern recognition depends precisely on exceeding human interpretive capacity. A weapon that forecloses what the law requires at the moment the law requires it is not a weapon that can be made compliant. It is a weapon that is incompatible by design.

The second proposition concerned the attribution framework of the Articles on State Responsibility. ARSIWA's effective control standard encounters a compliance deficit of categorical rather than remediable character when the operative decision layer is algorithmic. The standard presupposes a human decision-maker at the operative link of the causal chain — a human whose direction or control of the wrongful act can be

assessed under Articles 8 and 11. Tier 3 architecture eliminates that human at precisely that link. The causal chain does not break; it continues through the autonomous system's algorithmic processes to produce effects that no human authorized at the moment of their generation. This is not an evidentiary problem that better investigation will solve. It is a structural mismatch between the conceptual architecture of State responsibility law and the operational architecture of systems designed to eliminate the human from the causal chain. It cannot be closed by doctrinal extension operating on existing provisions.

The third proposition supplied the response to the deficit the first two create. Article 28(1) of the Rome Statute, reoriented to the deployment authorization stage through the constructive knowledge standard established in *Bemba Gombo*, supplies the legally coherent accountability locus that both deficiencies jointly require. The key move is temporal: accountability cannot be established at the operational phase because no human is present. It can be established at the deployment authorization stage because a human is present there — the commander who decides to deploy a system structurally incapable of IHL-compliant targeting. That commander, who should have known at the moment of authorization that the system's architecture would produce IHL-violating conduct, and who failed to conduct the Article 36 review that was the necessary and reasonable measure to prevent it, satisfies both elements of Article 28(1)(a). This is not new law. It is existing law applied at the correct temporal locus — the only locus where a legally cognizable human decision-maker exists.

Together, these three propositions establish the dissertation's thesis. IHL cannot govern Tier 3 autonomous cyber weapons at the moment of targeting because no human is there. The law must therefore be applied at the only moment where a human remains: the moment of design and the moment of authorization.

II. The Institutional Consequence

From the doctrinal confirmation of the hypothesis, two institutional consequences follow with legal necessity.

The first is Human Security by Design. If IHL cannot operate at the moment of targeting, its obligations must be relocated to the design phase — translated into technical specifications that a system must satisfy before it can be lawfully deployed. Those specifications are not new law. They are what existing law already requires, stated in language engineers can build and lawyers can verify: a kill-switch override operationalizing Article 57(2)(b), geofencing operationalizing Article 48, temporal deactivation logic operationalizing Article 57(1), and tamper-proof audit logging operationalizing Article 28(1) of the Rome Statute. A system that satisfies all four has been redesigned so as to no longer structurally eliminate the human at the operative moment. It has crossed the threshold from prohibited Tier 3 to permissible Tier 2. A system that cannot satisfy them cannot be lawfully deployed. That is not a policy preference. It is what the structural incompatibility established in Chapters 2, 3 and 4 demands.

The second is the IAWCO. A normative framework without an institution to verify and enforce it is an aspiration, not law. The International Autonomous Weapons Control Organization proposed in Chapter 6 provides the pre-deployment certification regime, the Technical Secretariat verification architecture, and the graduated enforcement mechanism through which Human Security by Design becomes binding international law rather than an academic proposal. The three-instrument sequencing — a binding CCW Protocol in the short term, the IAWCO Convention in the medium term, and a Security Council resolution as the normative catalyst — represents the only architecture whose legal authority exists now under existing Charter provisions without amendment of any international instrument.

III. Policy Recommendations

The governance architecture recommended by this dissertation comprises three sequenced instruments. In the short term (zero to two years), a binding CCW Protocol establishes the per se prohibition on Tier 3 deployment against objects protected under AP I Articles 12, 54(2), and 56(1). In the medium term (two to seven years), the IAWCO Convention, modelled on the CWC institutional architecture, serves as the primary normative instrument, extending governance to the full weapons lifecycle. Accompanying both instruments, a Security Council resolution under Chapter VII — or, where permanent member consensus is unavailable, a Chapter VI recommendation — functions as the normative catalyst that establishes the principle of international legal constraint before treaty negotiation is complete. These three instruments are not alternatives; they are a sequence. The CCW Protocol is the floor. The IAWCO Convention is the architecture. The Security Council resolution is the accelerant.

High Contracting Parties to Additional Protocol I bear, as a matter of binding treaty obligation, a non-derogable duty to conduct mandatory legal reviews pursuant to Article 36 in respect of all autonomous cyber weapon systems prior to their study, development, acquisition, or adoption, a duty whose normative foundation is the cardinal principle, codified in Article 35(1) of the Protocol, that the right of parties to an armed conflict to choose methods or means of warfare is not unlimited.[437] That obligation admits of no exception by reference to the novelty of the technological modality: the Protocol's text directs the review obligation to "any new weapon, means or method of warfare," terms which, on their ordinary meaning within the interpretive

437 AP I, art. 35(1) ("In any armed conflict, the right of the Parties to the conflict to choose methods or means of warfare is not unlimited"); *id.* art. 36 ("In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party"); ICRC Article 36 Guide, *supra* note 248, at 1–3.

framework of Article 31(1) VCLT, encompass any instrument through which a party to an armed conflict seeks to produce coercive or destructive effects upon the adversary, irrespective of whether those effects are generated through kinetic or digital delivery mechanisms.[438] That obligation extends to systems under development and to pre-positioned systems already deployed. States that have pre-positioned Tier 3 systems against critical civilian infrastructure bear a continuing obligation under Article 57(2)(b) AP I, and in the countermeasures framework of ARSIWA Articles 49–54 subject to the absolute prohibitions of Article 50(1)(c) and (d),[439] to neutralize or recall those systems to the extent technically feasible.

The Security Council retains full Chapter VII authority to address autonomous cyber weapons as threats to international peace and security under Charter Articles 39, 41, and 42,[440] and should establish a working group with a mandate to develop a binding framework prohibiting Tier 3 autonomous operation against critical civilian infrastructure and requiring States to declare and subject to international verification all Tier 3 cyber weapon programmes. The ICRC's most urgent contribution is a formal institutional acknowledgment — equivalent in weight to its 2019 position on autonomous weapon systems and its 2026 IHL guidance — that AP I's precautionary architecture is architecturally

438 Vienna Convention on the Law of Treaties, *supra* note 89, art. 31(1) ("A treaty shall be interpreted in good faith in accordance with the ordinary meaning to be given to the terms of the treaty in their context and in the light of its object and purpose"); ICRC Article 36 Guide, *supra* note 248, at 1 ("The requirement that the legality of all new weapons, means and methods of warfare be systematically assessed is arguably one that applies to all States, regardless of whether or not they are party to Additional Protocol I.").

439 AP I, arts. 57(2)(b), 57(1); ARSIWA, *supra* note 21, arts. 49–54 (art. 49(1): countermeasures may only be taken "against a State which is responsible for an internationally wrongful act in order to induce that State to comply with its obligations"; art. 49(3): countermeasures "shall, as far as possible, be taken in such a way as to permit the resumption of performance of the obligations in question"; art. 50(1)(c)–(d): countermeasures shall not affect "obligations of a humanitarian character prohibiting reprisals" nor "other obligations under peremptory norms of general international law"; art. 51: countermeasures "must be commensurate with the injury suffered, taking into account the gravity of the internationally wrongful act and the rights in question"; art. 52(1): an injured State shall call upon the responsible State to fulfil its obligations before taking countermeasures; art. 53: countermeasures "shall be terminated as soon as the responsible State has complied with its obligations under part two in relation to the internationally wrongful act").

440 U.N. Charter, *supra* note 35, arts. 39, 41, 42; CWC, *supra* note 400, art. VIII(2) (establishing the OPCW as the institutional implementation body, providing the structural precedent for a Technical Secretariat with weapons review authority operating under a binding convention on weapons of mass destruction).

incapable of governing Tier 3 autonomous cyber weapons in operation, and that compliance must accordingly be assessed at the design and authorization stage.

Technology companies developing machine-learning systems with foreseeable military applications bear a due diligence obligation under the UN Guiding Principles on Business and Human Rights to conduct mandatory IHL impact assessments prior to entering defense contracts, to disclose foreseeable weaponization risks to the IAWCO Technical Secretariat, and to refrain from the knowing supply of commercial AI infrastructure for Tier 3 autonomous cyber weapon development. Civil society organizations should direct their norm-entrepreneurial capacity, demonstrated through the Ottawa and Oslo Processes in which civil society stigmatized anti-personnel mines and cluster munitions and drove adoption of binding prohibition instruments outside the United Nations framework,[441] toward the specific operational architecture of Tier 3 deployment: pre-positioning, dormancy, and condition-based activation without human oversight at the operative moment of harmful effect.

441 Martha Finnemore & Kathryn Sikkink, *International Norm Dynamics and Political Change*, 52 Int'l Org. 887, 896–99 (1998); Ottawa Treaty pmbl., *supra* note 418; Convention on Cluster Munitions pmbl. ¶ 16, May 30, 2008, 2688 U.N.T.S. 39; Casey-Maslen, Ottawa Convention Commentary, *supra* note 418, at 2.

IV. Avenues for Further Scholarship

This dissertation identifies four areas in which further legal scholarship is required.

First, empirical testing of the IHL compliance capacity of specific autonomous cyber weapon architectures - examining, through controlled technical analysis, whether and under what conditions Tier 2 systems can satisfy the verification and proportionality obligations of AP I — would supply the factual foundation for the Article 36 review criteria proposed in Chapter 5; particular attention should be paid to Schmitt and Thurnher's analysis of the feasibility standard and the proportionality "military advantage" algorithm problem,[442] and, as a separate but related inquiry, to the question of how Article 50(1) AP I's presumption of civilian status in case of doubt as to persons, and Article 52(3) AP I's presumption of civilian use in case of doubt as to objects, translate into operationally enforceable constraints on autonomous targeting architectures.[443]

Second, a comparative analysis of domestic weapons review frameworks beyond the jurisdictions examined here, including the practice of France, Germany, and the United Kingdom, is necessary to identify workable consensus criteria for the IAWCO declaration regime.

Third, the evidentiary standard governing the trigger of the challenge inspection mechanism under Article 7(2)(c) of the proposed IAWCO Convention warrants dedicated doctrinal analysis: the CWC's challenge inspection procedure has never been formally invoked in nearly three

442 Michael N. Schmitt & Jeffrey S. Thurnher, *"Out of the Loop": Autonomous Weapon Systems and the Law of Armed Conflict*, 4 Harv. Nat'l Sec. J. 231, 255–57, 260–61 (2013) [hereinafter Schmitt & Thurnher, *Out of the Loop*]; *see also* 1987 ICRC AP Commentary, *supra* note 18, ¶¶ 2208, 2210 (confirming that proportionality assessments "must above all be a question of common sense and good faith for military commanders.

443 Schmitt & Thurnher, *Out of the Loop*, *supra* note 442, at 257, 263 (analyzing the proportionality requirement as applied to autonomous weapon systems, noting that "for the immediate future . . . the actual proportionality decision will continue to be made by humans, by deciding to launch the system into a particular environment, by deciding on how to preprogram the system, or by revising the engagement criteria remotely," and that the art. 57 feasibility standard requires that autonomous weapon systems "not be permitted to engage a target in situations in which a reasonable attacker would elect not to attack because of the degree of uncertainty").

decades of operation, and establishing an evidence threshold that renders the mechanism credibly triggerable without transforming it into an instrument of geopolitical competition is the most consequential open question in the Convention's verification architecture.

Fourth, the quantum computing implications for autonomous cyber weapon architecture, specifically, whether quantum–enabled cryptographic capabilities alter the attribution resistance characteristics analyzed in Chapter 1, represent a frontier issue requiring dedicated technical–legal engagement that the existing IHL literature has not yet addressed.

V. The Imperative of Preventive Governance

The Volt Typhoon intrusions, documented in the joint advisory issued by CISA, NSA, FBI, and partner agencies on 7 February 2024, confirm that PRC State-sponsored actors have maintained persistent access to United States critical infrastructure, spanning Communications, Energy, Transportation, and Water and Wastewater Systems sectors, for periods exceeding five years, and have demonstrated a confirmed capability to move laterally from information technology networks to operational technology control systems with the capacity to disrupt energy distribution, water pressure management, and transportation coordination.[444]

The operational architecture disclosed in those advisories- pre-positioning, dormancy, condition-based activation, and

444 CISA, NSA, FBI et al., *Joint Cybersecurity Advisory AA24-038A, PRC State-Sponsored Actors Compromise and Maintain Persistent Access to U.S. Critical Infrastructure* 1, 7–8 (Feb. 7, 2024), https://www.cisa.gov/news-events/cybersecurity-advisories/aa24-038a (last visited Apr. 16, 2026) ("the U.S. authoring agencies assess with high confidence that Volt Typhoon actors are pre-positioning themselves on IT networks to enable lateral movement to OT assets to disrupt functions"; actors maintained access "for periods exceeding five years"; "in one confirmed compromise, Volt Typhoon actors moved laterally to a control system and were positioned to move to a second control system"); *see also* NSA, CISA, FBI et al., *Joint Cybersecurity Advisory AA23-144A, People's Republic of China State-Sponsored Cyber Actor Living off the Land to Evade Detection* 1–2, 5–8 (May 24, 2023), https://media.defense.gov/2023/May/24/2003229517/-1/-1/0/CSA_PRC_State_Sponsored_Cyber_Living_off_the_Land_v1.1.PDF (last visited Apr. 16, 2026) ("one of the actor's primary tactics, techniques, and procedures (TTPs) is living off the land, which uses built-in network administration tools to perform their objectives"; confirmed NTDS.dit exfiltration via Volume Shadow Copy and persistent domain compromise while "blending in with normal Windows system and network activities").

the deliberate elimination of the human operator at the moment of harmful effect, constitutes the paradigmatic anatomy of an autonomous cyber weapon. It is that anatomy, not any abstract theoretical category, that this dissertation's proposed regulatory framework is designed to govern, and it is the documented presence of that anatomy in active deployment against protected civilian infrastructure that makes the framework not merely academically defensible but operationally necessary.

The window for preventive governance remains open. The operational record surveyed in this dissertation establishes with precision both the urgency of closing it and the doctrinal architecture required to do so. The structural analysis set out in the preceding chapters has established what that institutional design must be, why existing doctrine cannot supply it, and why the legal authority to establish it through a General Assembly-convened diplomatic conference exists now, without amendment of any existing international instrument, under Charter Article 13(1)(a). The evolutionary development of AP I Article 36's weapons review obligation to encompass multilateral institutional oversight constitutes the dissertation's *de lege ferenda* contribution, grounded in but not yet confirmed as *lex lata*.

VI. The Answer

The Introduction asked whether international humanitarian law, as currently constructed, is structurally incompatible with Tier 3 autonomous cyber weapons — not because its rules are wrong, but because its compliance architecture presupposes a human being present at the moment of lethal targeting, a precondition Tier 3 systems eliminate by design.

The answer is yes. Confirmed. Without qualification.

IHL's rules are not wrong. They are, in the context of kinetic warfare with human combatants, among the most sophisticated normative achievements in the history of international law. They are incapable of governing Tier 3 autonomous cyber weapons for one reason only: they were constructed for a world in which a human being stands between the weapon and its effect. Tier 3 systems have been engineered to remove that human. They remove not merely a compliance mechanism but the precondition of compliance itself.

The dissertation's six chapters form a single answer to that single question. Chapter 1 defined the weapon that removes the human. Chapter 2 demonstrated what the *jus contra bellum* cannot accommodate when that human is absent at the moment of force. Chapter 3 demonstrated what the *jus in bello* cannot accommodate when that same human is absent at the moment of attack. Chapter 4 demonstrated the accountability consequences of that dual structural failure. Chapter 5 provided the normative answer: Human Security by Design. Chapter 6 built the institution that makes that answer enforceable: the IAWCO.

A legal framework whose compliance architecture depends on a human being present at the moment of its operation cannot function as a legal constraint when the weapon it governs has been designed to eliminate that human. That is this dissertation's finding. Human Security by

Design is its consequence: where law cannot operate at the moment of targeting, it must operate at the moment of design.

That is not a policy proposal. That is what international humanitarian law, properly understood, already requires.

Syria knew what weapons do to civilian life. The law has always known what it requires in response: that those who design weapons capable of eliminating the human it presupposes must be held to account at the moment of design — before deployment, before activation, before the law finds no one left to speak to. That obligation already exists. This dissertation has only given it a name.

DOCTORAL DISSERTATION

Autonomous Cyber Weapons and International Humanitarian Law: New Protective Architectures for Human Security

Hassan Khoder, LL.M.

SUPPLEMENTARY ANNEX

Annex I: Draft Convention on the Control of Autonomous Cyber Weapons

ANNEX I

Draft Convention on the Control of Autonomous Cyber Weapons

(Model Text, Prepared for Academic and Policy Purposes)

Introductory Note

This Model Convention constitutes a scholarly contribution to the academic and policy debate on the international legal governance of autonomous cyber weapons, prepared as an annex to a doctoral dissertation submitted at the University of Cologne Faculty of Law in 2026. Its function is to translate the normative framework advanced in Chapters 5 and 6 of the dissertation into the formal legal language of a binding multilateral treaty, structured for adoption by States through the treaty-making procedures of the United Nations General Assembly.

The instrument draws upon the structural precedents established by the Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on Their Destruction, Jan. 13, 1993, 1974 U.N.T.S. 45 (entered into force Apr. 29, 1997) (CWC); the Rome Statute of the International Criminal Court, July 17, 1998, 2187 U.N.T.S. 90 (entered into force July 1, 2002); the Treaty on the Non-Proliferation of Nuclear Weapons, July 1, 1968, 729 U.N.T.S. 161 (entered into force Mar. 5, 1970) (NPT); the Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on their Destruction, Sept. 18, 1997, 2056 U.N.T.S. 211 (entered into force Mar. 1, 1999) (Ottawa Convention); and the institutional precedent of the Preparatory Commission for the Comprehensive Nuclear-Test-Ban Treaty Organization, established by Resolution 1 of the CTBT Signing

Conference, Sept. 24, 1996, as the model for an international organization exercising its institutional mandate in the pre-entry-into-force period pending the ratification threshold established in Article 18 of this Convention. Where treaty language is adapted from those instruments, this is identified in the Negotiating Commentary following each Article.

Each Article is followed by a Negotiating Commentary explaining the rationale for the specific language adopted, the treaty precedent upon which it draws, and the interpretive ambiguities the provision is designed to forestall. The Commentary is modelled upon the ICRC Commentaries to the Geneva Conventions and their Additional Protocols and is intended to function as a scholarly model commentary analogous in structure to those Commentaries, designed to anticipate interpretive questions that a future negotiating conference and treaty implementation bodies would need to resolve, and which such bodies may consult as a contribution to the academic literature informing the treaty's *travaux préparatoires* within the meaning of Article 32 of the Vienna Convention on the Law of Treaties, 23 May 1969, 1155 U.N.T.S. 331 (VCLT).

PREAMBLE

Reaffirming the purposes and principles of the Charter of the United Nations, and in particular the obligation of States under Article 2(4) to refrain in their international relations from the threat or use of force against the territorial integrity or political independence of any State, and the inherent right of individual or collective self-defence recognized in Article 51 thereof,

Recalling that the States Parties to the Geneva Conventions of 12 August 1949 have undertaken in common Article 1 thereof to respect and to ensure respect for those Conventions in all circumstances, and that this obligation imposes upon each High

Contracting Party a positive duty to ensure that the substantive protections of those instruments are operationally capable of being discharged, regardless of the means and methods by which an armed conflict is conducted,

Reaffirming the obligation established in AP I Article 57(2) to take all feasible precautions in the conduct of military operations with a view to avoiding, and in any event minimizing, incidental loss of civilian life, injury to civilians, and damage to civilian objects,

Reaffirming the principle enunciated in the Martens Clause, as expressed in the Preamble to the Hague Convention (IV) Respecting the Laws and Customs of War on Land of 18 October 1907 and codified in Article 1(2) of AP I, under which civilians and combatants remain, in all circumstances not covered by specific treaty provisions, under the protection of the principles of international law derived from established custom, the principles of humanity, and the dictates of public conscience,

Recognizing that the deployment of autonomous cyber weapons operating without real-time human control over individual targeting decisions creates the structural incapacity of human commanders to discharge the precautionary obligations established in AP I Article 57, a condition of operational precautionary inaccessibility, rendering reactive compliance assessment structurally inadequate as a protective mechanism for the persons and objects protected under international humanitarian law,

Recognizing further that autonomous cyber weapons directed against critical civilian infrastructure create cascade propagation risks that may, through the interdependencies of networked systems, extend harmful effects to protected persons and objects in States not party to any armed conflict and to civilian populations whose

protection cannot be anticipated at the time of deployment authorization,

Recognizing that the accountability gap arising from the temporal and epistemic disaggregation between deployment authorization and autonomous targeting decisions cannot be remedied through the application of existing legal doctrine alone, but demands preventive governance embedded at the development and acquisition stage, before autonomous decision-making capacity is deployed in conditions of armed conflict,

Concerned that AI-assisted semi-autonomous targeting systems, in which human commanders rely upon algorithmic targeting recommendations without meaningful capacity to verify the underlying analysis, generate accountability gaps that the existing law of armed conflict did not anticipate and that this Convention's preventive architecture is designed to address,

Determined to prevent, through legally binding international obligation, the erosion of the protections conferred by the principles of distinction, proportionality, and precaution in attack as they apply to the conduct of cyber operations involving autonomous targeting functions,

Convinced that an effective multilateral prohibition and control regime, established before autonomous cyber weapons achieve widespread operational deployment, represents the necessary and proportionate institutional response to the humanitarian challenge these systems present,

Have agreed as follows:

CHAPTER I
DEFINITIONS AND SCOPE

Article 1, Definitions

For the purposes of this Convention:

1. "Autonomous cyber weapon" means any cyber system, programme, algorithm, or combination thereof that, after deployment by human personnel, independently selects and produces coercive, disruptive, or destructive effects against targets through or by means of digital networks, without requiring real-time human authorization of individual targeting decisions, and that exhibits two or more of the following three operational characteristics:

(a) "Network propagation capacity": the capacity to disseminate, replicate, or extend its operational parameters beyond the initial target system through digital network connections without real-time human authorization of each individual propagation event;

(b) "Temporal persistence": the capacity to remain dormant within target systems or adjacent network infrastructure for a period exceeding twenty-four hours before activating primary operational effects;

(c) "Irreversible effects": the capacity to produce effects upon targeted systems that the deploying State cannot reverse, correct, or nullify following activation, including permanent data destruction, physical damage to hardware components through software-induced operational anomalies, or cascading infrastructure failure.

2. "Tier 3 autonomous cyber weapon" means an autonomous cyber weapon within the meaning of paragraph 1 that, once deployed, operates without any human capacity to monitor specific targeting decisions or to intervene, override, veto, or terminate individual engagement decisions prior to their execution.

3. "Critical civilian infrastructure" means systems and facilities whose disruption, destruction, or degradation of function would be expected to cause serious harm to civilian populations, including loss of life, injury, or the denial of essential services upon which civilian survival depends, including: electrical power generation and distribution systems; water treatment and distribution networks; healthcare facilities, hospital systems, and medical device networks; financial transaction processing systems; telecommunications infrastructure; transportation control and navigation systems; and food production and distribution control systems.

4. "Protected objects" means civilian objects within the meaning of Article 52(1) of AP I; objects indispensable to the survival of the civilian population within the meaning of Article 54 of AP I; works and installations containing dangerous forces within the meaning of Article 56 of AP I; cultural property within the meaning of Article 1 of the Convention for the Protection of Cultural Property in the Event of Armed Conflict of 14 May 1954; and medical units within the meaning of Articles 12 and 13 of AP I.

5. "IAWCO" means the International Autonomous Weapons Control Organization established under Article 14 of this Convention.

6. "Weapons review" means the legal assessment required under Article 36 of AP I to determine whether a new weapon, means, or method of warfare would, in some or all circumstances, be prohibited by applicable international law.

7. "Technical Secretariat" means the executive organ of IAWCO established under Article 15(3) of this Convention, including its Scientific and Technical Advisory Board.

8. "Behavioral testing" means the systematic assessment of an autonomous system's operational outputs under simulated conditions

designed to evaluate compliance with the prohibitions established in Article 3 and the precautionary standards established in Article 4(2).

Negotiating Commentary, Article 1: Article 1 translates the dissertation's four-characteristic analytical framework (Chapter 1, Section I.2) into binding definitional language, retaining three characteristics, network propagation, temporal persistence, and irreversible effects, within the Article 1(1) definitional threshold, and relocating the fourth, attribution resistance, to Article 3(6) as an aggravating circumstance, for the reasons given immediately below. The definition in paragraph 1 departs deliberately from the purely functional definitions of autonomous weapons that have circulated in CCW Group of Governmental Experts discussions by grounding the Convention's scope in operational architecture rather than targeting capability alone. The threshold of two or more characteristics prevents over-inclusion of systems that possess a single feature insufficient to generate the epistemic inaccessibility that is the structural basis of the prohibition, while ensuring that systems exhibiting the combination of characteristics that creates operational precautionary inaccessibility, most critically, temporal persistence combined with irreversible effects, or network propagation capacity combined with either temporal persistence or irreversible effects, fall unambiguously within scope. Attribution resistance has been relocated from the definitional threshold to Article 3(6), where it operates as an aggravating circumstance rather than a constitutive element of the prohibited category. This relocation prevents the perverse incentive that inclusion of attribution resistance in the definitional threshold would otherwise create: a State could not, by rendering its autonomous cyber weapon architecturally attributable, remove it from the Convention's categorical prohibition while retaining all operational characteristics, temporal persistence, network propagation, irreversible effects,

that generate the structural IHL compliance failure the Convention addresses.

Paragraph 2 incorporates the dissertation's three-tier classification. The designation "Tier 3" is a defined term rather than a descriptive label, to facilitate precise application of the categorical prohibition in Article 3 without requiring each application to recapitulate the definitional analysis. The expanded definition of "protected objects" in paragraph 4 adds medical units explicitly, beyond the categories enumerated in AP I Article 52, reflecting the humanitarian risk profile documented in Chapter 1's operational record: networked medical device systems deployed in hospital infrastructure constitute a class of protected object whose specific vulnerability to autonomous cyber weapons targeting existing definitional frameworks inadequately addressed. The phrase 'or combination thereof' in paragraph 1 closes a potential circumvention gap: sophisticated autonomous cyber weapons typically comprise multiple integrated components developed by distinct actors across the supply chain; the definition applies to the integrated system regardless of its compositional architecture, precluding a State from arguing that no single component falls within the definition.

The two-characteristic threshold governs the categorical prohibition in Article 3; systems exhibiting fewer than two characteristics remain subject to the weapons review and behavioral testing obligations of Articles 4 and 6. The definition excludes autonomous systems whose exclusive function is passive intelligence collection producing no coercive, disruptive, or destructive effects, as such systems do not engage AP I Article 49(1)'s attack threshold.

Article 2, Scope of Application

1. This Convention applies to autonomous cyber weapons as defined in Article 1, developed, produced, stockpiled, acquired, transferred, or deployed by States Parties or by natural or legal persons subject to the jurisdiction or control of States Parties, and to autonomous cyber systems approaching the definitional threshold of Article 1(1) that are subject to the weapons review and behavioral testing obligations established in Articles 4 and 6.

2. States Parties shall, through the national legislative measures required under Article 3(4), ensure that the prohibitions established in Article 3(1) apply as a matter of domestic criminal law to all persons within their jurisdiction or effective control, including members of non-State armed groups and private persons, regardless of whether those persons are acting at the direction of a State. Nothing in this Convention limits the application of customary international humanitarian law to non-State parties to armed conflicts, including the prohibitions on attacks against protected objects established in common Article 3 of the Geneva Conventions and in customary international law as reflected in the ICRC Customary IHL Study.

Negotiating Commentary, Article 2: The phrase 'natural or legal persons subject to the jurisdiction or control of States Parties' extends the Convention's personal scope to private actors, technology corporations, defense contractors, and cloud infrastructure providers, whose conduct in relation to autonomous cyber weapons is governed by States Parties' domestic implementing legislation required under Article 3(4). This extension responds to the dual-use supply chain challenge identified in Chapter 1 of the dissertation: the commercial AI sector's role in providing the components and infrastructure through which autonomous cyber weapons are developed and deployed cannot be excluded from the Convention's scope without

creating a governance gap sufficient to render the categorical prohibitions ineffective. The legal basis for this extension is established by Article VII(1) of the CWC, which requires States Parties to adopt national measures to implement their obligations and to extend those measures to persons on their territory, and CWC State practice has confirmed that such measures appropriately reach private sector actors.

Article 2(2) addresses the non-State actor gap that the Convention's State-Party framework would otherwise leave open. The technological trajectory of autonomous cyber weapons, decreasing cost, increasing availability of open-source AI architectures, means that non-State armed groups will be capable of deploying systems meeting the Article 1(1) definition within the Convention's operational lifetime. The reference to the ICRC Customary IHL Study in Article 2(2) reflects Jean-Marie Henckaerts and Louise Doswald-Beck's authoritative compilation, which documents the customary character of those prohibitions as binding on all parties to armed conflict.[445] The reference ensures that the Convention's humanitarian protections are not limited to States Parties to AP I but extend, through the domestic implementing legislation required under Article 3(4), to all parties to armed conflicts in which the systems governed by this Convention may be employed.

The reference to systems "approaching the definitional threshold" ensures that the weapons review and behavioral testing obligations apply at the pre-deployment stage, operationalizing the Convention's preventive architecture at the point where regulatory intervention retains the greatest capacity to prevent harm.

445 Jean-Marie Henckaerts & Louise Doswald-Beck, Customary International Humanitarian Law (Cambridge Univ. Press 2005); *see also* ICRC, *Customary IHL Database*, https://ihl-databases.icrc.org/customary-ihl (last visited May 28, 2026).

CHAPTER II
PROHIBITIONS

Article 3, General Prohibitions

1. Each State Party undertakes never under any circumstances to: (a) develop, produce, stockpile, or otherwise acquire Tier 3 autonomous cyber weapons directed, or designed to be directed, against protected objects or critical civilian infrastructure;

(b) transfer, directly or indirectly, Tier 3 autonomous cyber weapons directed, or designed to be directed, against protected objects or critical civilian infrastructure, to any recipient whatsoever;

(c) deploy Tier 3 autonomous cyber weapons against protected objects or critical civilian infrastructure; and

(d) assist, encourage, or induce, in any way, any natural or legal person or State to engage in any activity prohibited to States Parties under this Convention.

2. The prohibitions in paragraph 1 apply regardless of whether the protected objects or critical civilian infrastructure targeted are located in the territory of a State Party, a State not party to this Convention, or a territory under the control of a non-State party to an armed conflict.

3. Each State Party undertakes to destroy or otherwise render inoperable all Tier 3 autonomous cyber weapons within its possession, stockpile, custody, or control that are directed, or designed to be directed, against protected objects or critical civilian infrastructure, in accordance with the procedures to be established in the Verification Annex, within twenty-four months of its initial declaration submitted under Article 5. For the purposes of this paragraph, destruction of an autonomous cyber weapon means the irreversible deletion of all copies of the system's operational code, machine learning

weights, activation logic, and associated command-and-control infrastructure from all servers, storage devices, and network environments under the State Party's control, verified through the declaration and inspection procedures established in the Verification Annex. Relocation of operational code to air-gapped or otherwise isolated storage facilities does not constitute destruction or rendering inoperable within the meaning of this paragraph.

4. Each State Party shall adopt the national legislative, regulatory, and administrative measures necessary to implement the prohibitions in paragraph 1 and the destruction obligation in paragraph 3, including the establishment of national criminal offences applicable to natural and legal persons under its jurisdiction or control.

5. The prohibition in paragraph 1 applies without exception and cannot be suspended, derogated from, or qualified on the basis of any claim of self-defence under Article 51 of the Charter of the United Nations, any Security Council authorization under Chapter VII of the Charter, or any claim of military necessity.

Negotiating Commentary, Article 3 (Supplemental): The categorical character of the prohibition in paragraph 1 derives from the structural incompatibility of Tier 3 autonomous cyber weapons directed against protected objects with the precautionary obligations of AP I Article 57, which the right of self-defence and the doctrine of military necessity do not displace and cannot override. Paragraph 5's exclusion of Chapter VII authorization from the grounds on which the prohibition may be qualified does not contradict Article 103 of the UN Charter, which establishes the supremacy of Charter obligations over conflicting treaty obligations but does not extend to obligations of a jus cogens character; a Security Council resolution authorizing conduct that would itself violate a peremptory norm is void and without legal effect to that extent. The drafters take the position that

the prohibition on directing Tier 3 autonomous cyber weapons against protected objects and critical civilian infrastructure reflects the peremptory character of the underlying IHL protections it operationalizes, placing paragraph 5 outside, rather than in conflict with, Article 103's ordering rule.

6. For the purposes of this Convention, the presence of attribution-resistance features in a Tier 3 autonomous cyber weapon constitutes an aggravating circumstance in any determination of non-compliance under Article 12 and in any assessment of individual criminal responsibility under Article 13. "Attribution resistance" means design or operational features that systematically impede identification of the deploying State or the reconstruction of the causal chain connecting a specific activation event to the human authorization decisions that initiated deployment.

Negotiating Commentary, Article 3: The formulation 'never under any circumstances' is drawn verbatim from Article I(1) of the CWC and Article 1 of the Ottawa Convention and establishes an absolute prohibition admitting no derogation. This absolute framing is not a political judgment of convenience but the legally compelled consequence of the dissertation's structural incompatibility thesis: where *verification impossibility* forecloses the capacity of a human commander to discharge the precautionary obligations mandated by AP I Article 57 with respect to Tier 3 autonomous cyber weapons directed against protected objects, there exist no circumstances in which a lawful targeting decision can be made, and consequently no circumstances in which deployment of such weapons against such objects can be legally authorized.

The phrase "directed or designed to be directed" covers both systems actually targeted at protected objects at the time of prohibition and systems whose design parameters make the targeting of protected

objects foreseeable under operationally realistic deployment conditions, preventing circumvention through post-deployment configuration alterations. This anticipatory framing tracks the structure of AP I Article 35(2), which prohibits weapons calculated to cause unnecessary suffering, and AP I Article 51(4)(c), which prohibits area attacks of a nature to strike civilians indiscriminately, both formulations operating on the basis of a system's design characteristics rather than exclusive proof of specific harmful effects.

The parallel to Common Article 3 of the Geneva Conventions, which the ICJ confirmed in *Nicaragua*, ¶ 218, applies in all circumstances regardless of the legal character of the conflict and regardless of any claim of military necessity, is exact and intentional.

CHAPTER III
WEAPONS REVIEW, DECLARATION AND BEHAVIORAL TESTING

Article 4, Mandatory Weapons Review

1. Each State Party shall establish and maintain a national legal review authority responsible for conducting weapons reviews of all autonomous cyber systems developed, produced, or acquired for military purposes, to determine whether such systems constitute Tier 3 autonomous cyber weapons within the meaning of Article 1(2).

2. Weapons reviews conducted under paragraph 1 shall assess:

(a) whether the system, as designed and when operating within design parameters, is capable of selecting and producing coercive, disruptive, or destructive effects against protected objects or critical civilian infrastructure;

(b) whether the system's targeting architecture provides adequate human control over individual engagement decisions to satisfy the precautionary obligations established in AP I Article 57, including

the capacity to verify target status and conduct proportionality assessments at the time of each attack;

(c) whether the system's network propagation characteristics create foreseeable risk of cascade effects extending to protected objects or critical civilian infrastructure under conditions of distribution shift, including changes in network topology occurring during any dormancy period; and

(d) whether the system's temporal persistence characteristics would render the precautionary assessments conducted at deployment authorization inapplicable to attacks occurring during the dormancy period, by reason of material changes in target status, network topology, the cessation of hostilities, the conclusion of a ceasefire agreement, or any other legal development that would render the original military objective invalid under applicable international law.

3. Weapons reviews shall be conducted by legally qualified personnel possessing the technical expertise necessary to assess the algorithmic architecture of the system under review. States Parties shall consult the Scientific and Technical Advisory Board established under Article 15(4) where a review raises questions of technical or legal interpretation of general application.

4. Each State Party shall transmit a summary of each weapons review determination to the Technical Secretariat within sixty days of completion, including the system's assessed classification, the principal factors considered, and the outcome of the assessment.

5. This Article does not limit any State Party's obligations under Article 36 of AP I of 8 June 1977; the weapons review requirements established by this Convention are additional to, and not substitutes for, the obligation established therein.

6. Where a weapon system below the definitional threshold of Article 1(2) incorporates machine learning components that generate targeting recommendations upon which human operators foreseeably rely without independent capacity to verify the underlying algorithmic analysis, the weapons review conducted under paragraph 1 shall assess whether the human operator's effective capacity to exercise meaningful control over individual targeting decisions satisfies the precautionary obligations of AP I Article 57. A system whose design makes independent verification of algorithmic targeting recommendations operationally infeasible shall be treated, for the purposes of this Article, as equivalent to a Tier 3 autonomous cyber weapon.

Negotiating Commentary, Article 4: Article 4 operationalizes the preventive governance architecture that constitutes the dissertation's principal *de lege ferenda* contribution: the proposition that the prohibitions of international humanitarian law must, in the case of autonomous cyber weapons, be enforced at the development and acquisition stage rather than solely through post-hoc accountability mechanisms applied after harmful effects have materialized.

Paragraphs 2(c) and 2(d) are the normatively consequential provisions. They require weapons review authorities to assess not only whether a system complies with IHL at the moment of deployment, but whether IHL compliance can be maintained across the temporal gap between deployment authorization and autonomous activation, what this dissertation designates the temporal and epistemic disaggregation between deployment authorization and autonomous targeting decisions. Standard weapons review procedures developed for kinetic weapons systems are structurally unsuited to autonomous cyber weapons that remain dormant in target networks for extended periods before activation. An assessment conducted at deployment authorization may accurately describe the system's compliance characteristics as of that date; it cannot account for changes in target

status, network topology, or the law applicable to the conflict that occur during the dormancy period. Paragraphs 2(c) and 2(d) require the weapons review authority to address this gap explicitly, binding it to assess the system's compliance characteristics not only at deployment but across the reasonably foreseeable range of operational conditions it will encounter prior to activation.

Article 4(5) preserves the relationship with the AP I Article 36 obligation: this Convention's weapons review requirements are additional obligations, not substitutes for the customary obligation to review new means and methods of warfare.

Article 4(6) addresses the accountability gap that AI-assisted semi-autonomous targeting systems generate, a challenge that the CCW Group of Governmental Experts has failed to engage with doctrinal precision and that the operational record of AI-assisted targeting in recent armed conflicts has rendered urgent. The provision captures systems that fall formally below the Article 1(2) Tier 3 threshold, because a human nominally authorizes each engagement, but which generate equivalent accountability gaps in practice, because the human operator lacks the technical capacity to assess whether the algorithmic recommendation is correct. The standard 'operationally infeasible' is drawn from AP I Article 57's 'feasible precautions' formulation: feasibility is an operational concept, not a theoretical one, and a verification capacity that exists in principle but cannot be exercised under the time pressure and information conditions of actual operations does not satisfy the AP I standard.

Article 5, Declarations

1. Not later than ninety days after the entry into force of this Convention for a State Party, that State Party shall submit to the Technical Secretariat an initial declaration specifying:

(a) whether the State Party possesses any Tier 3 autonomous cyber weapons within the meaning of Article 1(2);

(b) the locations at which such weapons are developed, produced, maintained, or stored;

(c) the general technical characteristics of each declared weapon, including the categories of network environments for which the system is designed, the type of condition-based activation mechanism, and the functional categories of target system types — for example, industrial control systems, power distribution networks, or financial transaction infrastructure, for which the weapon is operationally configured, without disclosure of specific identified target locations, networks, or entities; and

(d) all active deployment operations involving declared weapons at the time of the initial declaration, specifying the date of each deployment authorization and the general geographic scope of deployment.

2. Each State Party shall submit updated declarations to the Technical Secretariat on an annual basis by 15 February of each calendar year, covering the preceding calendar year.

3. Declarations submitted under this Article shall be treated as confidential by the Technical Secretariat and shall not be disclosed to other States Parties except with the express consent of the declaring State Party, or pursuant to a compliance determination by the Executive Council under Article 12.

Negotiating Commentary, Article 5: The declaration regime follows the structure of CWC Article III, which requires States Parties to declare existing stockpiles, production and processing facilities, and historical transfer records. Article 5(1)(d) extends this framework to active deployment operations, a provision without direct precedent in existing multilateral arms control instruments, but one necessitated by the temporal persistence characteristic of

autonomous cyber weapons: a system deployed before the Convention's entry into force but still dormant and active at that date must be declared and its continued operation addressed through the destruction obligation under Article 3(3), since an undeclared dormant weapon constitutes precisely the invisible threat the Convention's preventive architecture exists to address.

The confidentiality provision in paragraph 3 reflects States' legitimate security interests in the protection of intelligence sources and methods, including the existence and parameters of deployed cyber capabilities. The balanced approach, confidentiality as default, override available through the compliance mechanism under Article 12, follows the model of the IAEA Additional Protocol (INFCIRC/540), which protects sensitive information while preserving institutional oversight.

Article 6, Behavioral Testing and Compliance Certification

The certification regime established in this Article operates as an instrument of reclassification, not of authorization. A system submitted for behavioral testing under paragraph 1 is presumed to satisfy the Tier 3 definition of Article 1(1) until the Technical Secretariat affirmatively certifies that the embedded restraints introduced under the Human Security by Design framework have reduced the number of Article 1(1) operational characteristics the system exhibits below the two-characteristic threshold established in Article 1(1), addressing network propagation capacity, temporal persistence, and irreversible effects as applicable — together with the attribution-resistance feature addressed in Article 3(6). Systems that satisfy the certification criteria have thereby been redesigned below the Tier 3 threshold and may be deployed subject to the continuing supervision requirements of this Article. Systems that fail certification remain subject to the per se

deployment prohibition of Article 3 of this Convention, irrespective of any unilateral State determination to the contrary.

1. Each State Party shall submit autonomous cyber systems approaching the definitional threshold of Article 1(1) to standardized behavioral testing protocols established by the Technical Secretariat, in consultation with the Scientific and Technical Advisory Board, prior to any deployment of such systems.

2. The Technical Secretariat shall develop and maintain standardized behavioral testing protocols, including:

(a) testing of target identification accuracy under distribution-shift scenarios, using simulated conditions that differ from the system's training-data environment to assess performance degradation under operationally realistic conditions;

(b) testing of network propagation behavior in simulated critical civilian infrastructure environments, incorporating the interdependency patterns of electricity, water, healthcare, and telecommunications systems;

(c) testing of human override and termination capacity under adversarial conditions designed to simulate attempts to prevent or delay deactivation; and

(d) red-team adversarial testing to assess system behavior under conditions designed to induce targeting errors, unintended propagation, or activation responses not contemplated by the deployment authorization.

3. Systems that fail behavioral testing shall not be deployed pending redesign and successful re-testing. Systems that successfully complete behavioral testing shall receive a compliance certificate from the Technical Secretariat, subject to renewal at intervals not exceeding three years.

4. Compliance certification under this Article does not constitute a determination of lawfulness in any specific deployment context; compliance with the substantive prohibitions of this Convention and with the applicable rules of international humanitarian law in any specific operation remains the responsibility of the deploying State and, in respect of individual targeting decisions, of the commanders and operators bearing decision-making authority under applicable international humanitarian law and this Convention's Article 13.

Negotiating Commentary, Article 6: The four categories of behavioral testing in paragraph 2 are each directed at the humanitarian risks that the definitional characteristics in Article 1(1) generate, though the correspondence is functional rather than one-to-one. Paragraph 2(a) tests target identification accuracy under distribution-shift scenarios; this addresses the general autonomous targeting function described in Article 1(2) and the risk that a system certified for compliance in one network environment will misidentify protected objects when its operational context diverges from its training-data environment. Paragraph 2(b) tests network propagation behavior in simulated critical civilian infrastructure environments; this corresponds directly to the network propagation capacity characteristic in Article 1(1)(a) and assesses cascade risks that propagation into interdependent civilian systems generates. Paragraph 2(c) tests human override and termination capacity under adversarial conditions; this addresses the irreversible effects characteristic in Article 1(1)(c) by assessing whether effects can be halted before irreversibility is reached. Paragraph 2(d) subjects the system to red-team adversarial testing for unintended targeting, propagation, and activation responses; this addresses the attribution-resistance characteristic treated as an aggravating circumstance under Article 3(6) by simulating adversarial interference with the system's decision architecture.

The three-year certification renewal period addresses the specific risk that a system certified as compliant at deployment may develop non-compliant behaviors through operational machine-learning adaptation, acquiring characteristics during deployment that were absent at the time of testing. Article 6(4) is structurally necessary. Without it, a State might invoke a compliance certificate as a defense against an allegation of IHL violation in a specific operation, conflating the systemic characteristics assessed through behavioral testing with the context-dependent proportionality and precautionary assessments that only a human commander can perform at the time of each attack.

CHAPTER IV
VERIFICATION AND CONFIDENCE–BUILDING

Article 7, Verification Inspections

1. For the purpose of verifying compliance with this Convention, each State Party accepts the right of the Technical Secretariat to conduct inspections as provided in this Article and in accordance with the procedures to be established in the Verification Annex.

2. Each State Party shall permit the Technical Secretariat to conduct:

(a) routine inspections of declared development and production facilities, at intervals not exceeding eighteen months;

(b) short-notice inspections of any declared facility within seventy-two hours of written notification; and

(c) challenge inspections of any facility, whether declared or undeclared, in the territory of the inspected State Party, following a request by any other State Party which has concerns about whether a facility may be engaged in activities prohibited by this Convention.

3. Inspections conducted under paragraph 2 shall be subject to managed access procedures established in the Verification Annex, which shall provide for protection of legitimately classified information unconnected to the Convention's subject matter while ensuring that inspection teams obtain access sufficient to assess compliance with the prohibitions and obligations established in this Convention. The inspected State Party shall designate a national authority responsible for coordinating managed access arrangements with the inspection team prior to and during each inspection.

4. Inspection teams may include technical experts with expertise in artificial intelligence architecture, machine learning systems, and critical infrastructure engineering, whose participation shall be subject to the confidentiality obligations applicable to Technical Secretariat staff.

Negotiating Commentary, Article 7: The three-tier inspection regime follows the structure established in CWC Articles IV, V, VI, and IX. The challenge inspection mechanism in Article 7(2)(c) is the regime's most important institutional feature: it provides the capacity to investigate credible allegations of non-compliance without requiring the requesting State to produce evidence sufficient to initiate proceedings before the UN Security Council, a threshold that proved prohibitive in the context of chemical weapons use where a permanent member's veto power foreclosed Chapter VII enforcement action,[446] and that the Security Council's veto structure renders

446 *See, e.g.,* U.N. SCOR, 72d Sess., 7922d mtg. at 2, U.N. Doc. S/PV.7922 (Apr. 12, 2017) (Russian Federation veto of draft resolution S/2017/315, submitted by France, United Kingdom, and United States, condemning the April 4, 2017 Khan Shaykhun sarin attack and requiring Syrian government cooperation with the OPCW Fact-Finding Mission; 10 votes in favour, 2 against, 3 abstentions; draft resolution not adopted owing to the negative vote of a permanent member of the Council), https://undocs.org/en/S/PV.7922 (last visited July 24, 2026); U.N. SCOR, 72d Sess., 8105th mtg. at 3, U.N. Doc. S/PV.8105 (Nov. 16, 2017) (Russian Federation veto of draft resolution S/2017/962, submitted by France, Italy, Japan, Sweden, Ukraine, United Kingdom, and United States to renew the mandate of the OPCW-UN Joint Investigative Mechanism for Syria for one year; 11 votes in favour, 2 against, 2 abstentions; draft resolution not adopted owing to the negative vote of a permanent member of the Council; veto terminated the Mechanism following its attribution of the Khan Shaykhun sarin attack to the Syrian government), https://undocs.org/en/S/PV.8105 (last visited July 24, 2026).

structurally unavailable in many cases of State non-compliance. The managed access procedure in Article 7(3) is modelled on the IAEA's managed access framework, which has proven workable in practice: it enables the inspected State to protect legitimately classified information unconnected to the Convention's subject matter while permitting inspectors to obtain the access necessary to assess compliance.

Article 7(4) addresses a gap in existing arms control inspection regimes that is particularly acute in the autonomous weapons context. Verification of software-based autonomous systems requires the capacity to examine machine learning architectures, training data sets, and algorithmic decision structures, technical expertise that legal inspection teams alone cannot supply. Without this capacity, the inspection regime would possess the formal authority to verify compliance while lacking the substantive technical capacity to do so.

Article 8, Confidence–Building Measures

1. Each State Party shall report annually to the Technical Secretariat:

(a) any cyber operations conducted during the preceding year using autonomous systems incorporating machine learning components in targeting functions, including a description of the human control measures in place during each operation; and

(b) any activation event involving an autonomous cyber system that produces coercive, disruptive, or destructive effects on protected objects or critical civilian infrastructure not contemplated by the deployment authorization, within forty-eight hours of the State Party first becoming aware of such effects, with an obligation to submit a supplementary report to the Technical Secretariat within thirty days detailing the full scope, duration, and attribution analysis of the activation event.

2. The Technical Secretariat shall publish an annual public confidence-building report summarizing the information received under paragraph 1, omitting information the disclosure of which would compromise the legitimate national security interests of reporting States Parties.

Negotiating Commentary, Article 8: Article 8(1)(b) introduces a mandatory incident reporting obligation adapted from the Vienna Document 2011's military incident notification procedures to the specific risk profile of autonomous cyber weapons. The forty-eight-hour reporting window is technically justified: the deploying State possesses immediate access to deployment authorization records and system telemetry enabling preliminary assessment, and delay beyond forty-eight hours would materially impair the Technical Secretariat's capacity to investigate whether the incident reflects a broader pattern of non-compliance. The public annual report under Article 8(2) creates the transparency infrastructure necessary for civil society monitoring and independent academic assessment of compliance patterns, following the model of the Ottawa Convention's Article 7 transparency reports. The omission clause in paragraph 2 is calibrated to protect genuinely sensitive operational security information while preventing States from invoking national security interests to suppress information about activation events affecting protected objects, events whose public disclosure the Convention's humanitarian purpose demands.

CHAPTER V
COOPERATION, ASSISTANCE, AND RESEARCH

Article 9, Technical Assistance and Capacity–Building

1. States Parties in a position to do so shall, through IAWCO and through bilateral arrangements, provide technical and economic assistance to other States Parties in the implementation of this Convention, including assistance in:

(a) establishing, equipping, and training national legal review authorities under Article 4;

(b) developing national legislative frameworks implementing the prohibitions and obligations of this Convention; and

(c) building the technical capacity to conduct behavioral testing under Article 6.

Negotiating Commentary, Article 9: Article 9 follows the capacity-building model of Ottawa Convention Article 6. The positive obligation to provide assistance is structurally indispensable in the autonomous weapons context: the technical expertise required to conduct meaningful weapons review and behavioral testing of AI systems is concentrated in a small number of technologically advanced States, and without active capacity transfer, developing States will be structurally unable to implement the Convention's preventive governance architecture, creating a two-tier compliance regime in which the Convention's obligations bind, in practice, only States with pre-existing technical capacity. The requirement to channel assistance "through IAWCO" ensures that bilateral assistance programs are coordinated institutionally, preventing fragmentation and enabling the Technical Secretariat to identify and address systematic implementation gaps.

Article 10, Preferential Cooperation in Civilian Autonomous Systems

1. States Parties shall, to the fullest possible extent, exchange equipment, material, and scientific and technological information for the development and application of autonomous systems for exclusively peaceful civilian purposes, including in healthcare, agriculture, infrastructure management, and environmental monitoring.

2. No State Party shall impose restrictions on the transfer to other States Parties of technology for peaceful autonomous systems applications that are inconsistent with the object and purpose of this Convention.

Negotiating Commentary, Article 10: The phrase 'to the fullest possible extent' is drawn verbatim from CWC Article XI(1) and reflects the foundational arms control principle that restrictions on military applications must be balanced by affirmative facilitation of peaceful applications, the 'grand bargain' architecture that has proven, across the NPT and CWC regimes, to be the mechanism through which States that would otherwise resist restrictions on their military capabilities may be persuaded to join and remain in a multilateral control regime. Article 10(2)'s non-restriction obligation follows NPT Article IV(2) and is designed to prevent the Convention from serving as a legal pretext for technology transfer restrictions that have no connection to the autonomous cyber weapons the Convention exists to govern. The positive cooperation obligation in Article 10 constitutes one of the three compliance incentives in the enforcement architecture developed in Chapter 6 of the dissertation: access to cooperative benefits is contingent upon compliance, creating a positive inducement alongside the punitive mechanisms of Article 12.

Article 11, Scientific and Technical Cooperation and Review

1. States Parties shall promote research into:

(a) explainable artificial intelligence techniques capable of producing human-readable audit trails of autonomous system targeting decisions, enabling ex post verification of compliance with the principles of distinction and proportionality;

(b) technical standards for effective human override and deactivation systems that function under adversarial conditions, including conditions in which an adversary actively seeks to prevent deactivation of a deployed system; and

(c) forensic attribution methodologies applicable to autonomous cyber weapon activation events, enabling identification of the deploying State in cases where attribution-resistance features have been employed.

2. The Conference of States Parties shall, at each quinquennial Review Conference held under Article 15(5), assess whether advances in explainable artificial intelligence have materially altered the verification-impossibility problem upon which the categorical prohibition in Article 3 is premised. Any determination that such advances require amendment of the categorical prohibition shall be subject to the amendment procedure established in Article 19, requiring acceptance by a two-thirds majority of all States Parties.

Negotiating Commentary, Article 11: Article 11(2) is the Convention's most doctrinally innovative provision, and its inclusion reflects an obligation of intellectual honesty rather than political compromise. The categorical prohibition in Article 3 rests on the verification-impossibility thesis, a finding about the current State of a technology that continues to develop. It would be both intellectually indefensible and politically counterproductive for the Convention to treat that finding as a permanent, irrebuttable basis for prohibition

rather than as a contingent factual predicate that States Parties are entitled to reassess as the technology evolves. The quinquennial Review Conference mechanism provides the institutional vehicle for this reassessment, while the two-thirds majority requirement for any consequential amendment to Article 3 ensures that relaxation of the categorical prohibition requires genuine consensus. The research cooperation obligation in Article 11(1) simultaneously provides States Parties with a constructive pathway toward technologies that might eventually justify revision of the prohibition: the Convention invests in the science that could change its own scope, reflecting a genuinely evolving legal instrument rather than a static prohibition premised on a permanently negative assessment of human ingenuity. This self-correcting architecture distinguishes the Convention from the Ottawa Convention's absolute prohibition, which is premised on a categorical moral judgment rather than a contingent technological assessment and therefore admits no equivalent review mechanism.

CHAPTER VI
ENFORCEMENT AND COMPLIANCE

Article 12, Graduated Compliance Measures

1. In the event of serious concern about non-compliance by a State Party, the following graduated measures shall be available:

(a) The Technical Secretariat may request written clarification from the State Party concerned and shall report its findings to the Executive Council within thirty days of receipt of the response, or, in the absence of a response, within thirty days of the expiry of the response deadline;

(b) The Executive Council may, by a three-quarters majority of all members, recommend suspension of the rights and privileges of the

non-complying State Party under this Convention, pending remediation of the conduct constituting non-compliance;

(c) The Conference of States Parties may, by a two-thirds majority of all members, suspend the rights and privileges of the non-complying State Party, including voting rights and access to technical assistance and cooperation under Articles 9 and 10; and

(d) The Conference of States Parties may, by a two-thirds majority of all members, refer cases of serious non-compliance to the United Nations Security Council and the United Nations General Assembly, together with the Technical Secretariat's findings of fact and the Executive Council's recommendations.

2. The measures in paragraph 1 are without prejudice to any other remedies available under international law, including recourse to State responsibility before the ICJ and individual criminal accountability before the International Criminal Court in respect of war crimes arising from the deployment of autonomous cyber weapons.

3. In cases of urgent non-compliance threatening immediate harm to protected persons or objects, the Executive Council may, by a three-quarters majority of all members, recommend interim protective measures pending full compliance proceedings under paragraph 1, including the immediate suspension of active deployment operations found to involve Tier 3 autonomous cyber weapons directed against protected objects or critical civilian infrastructure.

Negotiating Commentary, Article 12: The graduated compliance mechanism follows the sequential structure established in CWC Article XII, escalating from Executive Council engagement to Conference of States Parties action and ultimately to Security Council referral. The critical innovation over CWC Article XII is Article 12(1)(c)'s suspension of access to the cooperation benefits established in Articles 9 and 10. This provision creates a compliance

incentive operating wholly independently of the Security Council enforcement architecture, a design reflecting the enforcement architecture developed in Chapter 6, Section IV.2 of this dissertation, where the structural inadequacy of Security Council–dependent enforcement in the autonomous weapons context is addressed through compliance incentives that operate through treaty constitutional authority rather than Charter Chapter VII. The suspension of cooperative benefits restructures the incentive environment so that non-compliance carries tangible material costs that a State cannot avoid by exercising the veto.

The voting thresholds, three-quarters of all members for Executive Council recommendations, two-thirds of all members for Conference decisions, prevent a minority from blocking legitimate compliance measures while ensuring that the mechanism cannot be activated by a simple majority against a State making a good-faith compliance effort. Article 12(3)'s interim protective measures provision has no precedent in existing multilateral arms control instruments and is justified by the operational specificity of the autonomous cyber weapon threat: an actively deployed autonomous cyber weapon can execute targeting decisions within hours of the Technical Secretariat becoming aware of potential non-compliance. The provision draws structural inspiration from Article 41 of the Statute of the International Court of Justice, June 26, 1945, 3 Bevans 1153, which authorizes the indication of provisional measures 'to preserve the respective rights of either party,' as interpreted by the Court in LaGrand (Ger. v. U.S.), Judgment, 2001 I.C.J. Rep. 466, ¶ 102 (June 27) to impose legally binding obligations pending final determination, and from Article 56 and Article 57(3)(c) of the Rome Statute, which authorize the Pre-Trial Chamber to take measures necessary to preserve evidence and to ensure the protection and privacy of victims and witnesses prior to the commencement of full proceedings.

Article 13, Individual Criminal Responsibility

1. Each State Party shall take the necessary legislative measures to ensure that the following constitute criminal offences under its national law, punishable by penalties reflecting the gravity of the conduct:

(a) ordering or authorizing the deployment of a Tier 3 autonomous cyber weapon against protected objects or critical civilian infrastructure in violation of Article 3, where the person giving the order knew, or, owing to the circumstances at the time of the authorization decision, "should have known", that the weapon system authorized for deployment constituted a Tier 3 autonomous cyber weapon within the meaning of Article 1(2) of this Convention;

(b) knowingly directing, overseeing, or aiding, abetting, or otherwise assisting in the development, production, or transfer of Tier 3 autonomous cyber weapons in violation of Article 3;

(c) willfully obstructing, intimidating, or providing materially false information to a Technical Secretariat inspection team exercising functions under Article 7.

2. Nothing in this Article affects the jurisdiction of the International Criminal Court in respect of war crimes arising from the deployment of autonomous cyber weapons, including under Article 8(2)(b)(i), (ii), (iv), and (xxiv) of the Rome Statute, or the exercise of universal jurisdiction by national courts in respect of such crimes under applicable customary international law.

3. Nothing in this Article shall be construed to require a State Party to establish criminal offences for conduct occurring prior to the entry into force of this Convention for that State Party. The nullum crimen sine lege principle, as a general principle of law within the meaning of Article 38(1)(c) of the ICJ Statute and as codified in Article 22 of the Rome Statute, applies to all criminal offences established under

this Article. States Parties are, however, encouraged to reflect, in the preamble to their implementing legislation, that the prohibitions established in this Convention codify and render specific obligations that derive, in part, from pre-existing customary international humanitarian law applicable to autonomous targeting systems.

Negotiating Commentary, Article 13: Article 13(1)(a) operationalizes the dissertation's reorientation of command responsibility to the deployment authorization stage, the central doctrinal contribution to individual accountability law advanced in Chapter 4 of the dissertation. Standard command responsibility doctrine, as expressed in Article 28 of the Rome Statute, requires that a military commander knew, or, owing to the circumstances at the time, "should have known", that forces under effective command and control were committing or about to commit crimes. The structural challenge posed by Tier 3 autonomous cyber weapons is that the epistemic inaccessibility of their targeting decisions severs the temporal and informational link between the commander's knowledge State and the weapon's effects: by the time a targeting decision is executed, neither real-time knowledge nor real-time intervention is possible. Article 13(1)(a) responds by relocating the epistemic standard to the deployment authorization decision, the moment at which a human commander retains the knowledge and capacity to determine the weapon's type and to assess whether it falls within the Convention's prohibition.

The "knew or "should have known"" formulation is drawn directly from Article 28(1)(a) of the Rome Statute, as articulated by Trial Chamber III in *Prosecutor v. Bemba Gombo*, ICC-01/05-01/08-3343, Trial Chamber III, Judgment Pursuant to Article 74 of the Statute, ¶¶ 170–173, 192–194, 432–434 (Mar. 21, 2016), which held that 'owing to the circumstances at the time, "should have known"' requires the commander to have taken active steps to obtain the

relevant information rather than remaining in deliberate or negligent ignorance. This formulation must, however, be read in light of *Prosecutor v. Bemba Gombo*, ICC-01/05-01/08 A, Judgment on the Appeal of Mr Jean-Pierre Bemba Gombo against Trial Chamber III's 'Judgment pursuant to Article 74 of the Statute,' ¶¶ 167–191 (June 8, 2018), in which the Appeals Chamber, by a 3–2 majority, reversed Bemba's conviction in its entirety, holding *inter alia* that the Trial Chamber had applied an insufficiently demanding standard of appellate deference in assessing whether Bemba had taken 'all necessary and reasonable measures' under Article 28(a)(ii). The Appeals Chamber's judgment has been criticized as having narrowed, rather than clarified, the practical reach of Article 28 command responsibility; see Sadat, *Prosecutor v Jean-Pierre Bemba Gombo*, 113 AJIL 353 (2019); SáCouto & Sellers, *The Bemba Appeals Judgment: Impunity for Sexual and Gender-Based Crimes*, 27 Wm. & Mary Bill Rts. J. 599 (2018–19). This dissertation's deployment-stage reorientation does not depend upon, and is not undermined by, the contested appellate disposition of the 'necessary and reasonable measures' limb in Bemba; it depends only upon the Trial Chamber's uncontested formulation of the *content* of the '"should have known"' constructive-knowledge standard, relocated in time to the deployment authorization decision rather than the operational phase. A commander who authorizes the deployment of a weapon without determining whether its type falls within the Convention's prohibition, a determination that the weapons review authority under Article 4 exists to enable, cannot shelter behind ignorance of the weapon's characteristics where the diligence of the position demands that such a determination be made.

Article 13(1)(b)'s offence of material contribution to prohibited development and production deliberately reaches private sector actors, engineers, software architects, and executives of defense technology

companies, tracking the structure of Article 25(3)(c) of the Rome Statute, which establishes individual criminal responsibility for a person who, for the purpose of facilitating the commission of a crime, aids, abets, or otherwise assists in its commission. A criminal offence regime that reaches only State officials while leaving commercial developers unaccountable creates a gap that States may exploit through privatization of prohibited activities.

Article 13(3) confronts directly the nullum crimen sine lege challenge that criminal law scholars will raise upon examining Article 13(1). The principle is absolute, no retroactive criminalization, and this Convention respects it without exception. The encouraged legislative preamble language in the final sentence of paragraph 3 serves a distinct purpose: it creates a narrative record, in domestic implementing legislation, that connects the Convention's prohibitions to pre-existing customary IHL obligations, supporting an argument before domestic courts that the underlying conduct, deploying weapons structurally incapable of complying with the precautionary obligations of AP I Article 57, was already prohibited under customary international law before the Convention entered into force for the State concerned.

CHAPTER VII
THE INTERNATIONAL AUTONOMOUS WEAPONS CONTROL ORGANIZATION

Article 14, Establishment and Legal Status of IAWCO

1. The States Parties hereby establish the International Autonomous Weapons Control Organization to achieve the object and purpose of this Convention, to ensure the implementation of its provisions including those for international verification of compliance, and to provide a forum for consultation and cooperation among States Parties.

2. All States Parties are ipso facto members of IAWCO. Membership in IAWCO and adherence to this Convention are co-extensive and inseverable.

3. The principal organs of IAWCO are: the Conference of States Parties; the Executive Council; and the Technical Secretariat, including its Scientific and Technical Advisory Board.

4. IAWCO shall possess international legal personality and such legal capacity as is necessary for the exercise of its functions and the fulfilment of its purposes, including the capacity to contract, to acquire and dispose of property, and to institute legal proceedings before competent national and international tribunals.

5. IAWCO and its staff shall enjoy such privileges and immunities as are necessary for the independent exercise of its functions, as set forth in a Headquarters Agreement between IAWCO and the State in which its headquarters are situated, to be concluded within one year of IAWCO's establishment.

Negotiating Commentary, Article 14: Article 14 follows the institutional model of the Organization for the Prohibition of Chemical Weapons, established under CWC Article VIII, and draws upon the

institutional design of the IAEA under its Statute, opened for signature on 26 October 1956. The ipso facto membership provision ensures that membership in IAWCO and adherence to the Convention are structurally inseparable: a State cannot remain a member of IAWCO while withdrawing from the Convention's substantive obligations, nor benefit from IAWCO's cooperative programs while evading the inspection and declaration obligations that constitute the institutional rationale for those programs.

The international legal personality provision in Article 14(4) reflects the principle affirmed in Reparation for Injuries Suffered in the Service of the United Nations, Advisory Opinion, 1949 I.C.J. Rep. 174, at 179 (Apr. 11) that an international organization established by States for defined purposes possesses, by virtue of the doctrine of implied powers, the international legal personality necessary to fulfill those purposes, a principle subsequently confirmed as applicable to specialized agencies and treaty bodies established within the United Nations system. Explicit statement in Article 14(4) removes all interpretive uncertainty and confirms IAWCO's capacity to conclude the headquarters agreement required under Article 14(5), to contract for inspection and behavioral testing services, and to initiate proceedings to protect the integrity of its institutional functions. The phrase 'to provide a forum for consultation and cooperation' reflects the deliberative function of IAWCO alongside its enforcement role: the Conference of States Parties is designed to enable States Parties to address emerging technical challenges and interpretive disputes through institutional dialogue before engaging the compliance mechanisms of Article 12.

Article 15, Institutional Architecture of IAWCO

1. The Conference of States Parties is the supreme organ of IAWCO. It shall consider all questions within the scope of this Convention, including questions of implementation and compliance; adopt the programme and budget of IAWCO by consensus or, failing consensus, by a two-thirds majority of members present and voting; elect the members of the Executive Council; and appoint the Director-General of the Technical Secretariat. Substantive decisions shall be taken by a two-thirds majority of members present and voting, provided that a majority of all States Parties are present and voting.

2. The Executive Council is the executive organ of IAWCO, responsible for supervising the activities of the Technical Secretariat, ensuring the implementation of the decisions of the Conference of States Parties, and initiating compliance measures under Article 12. It shall consist of twenty-one members elected by the Conference on the basis of equitable geographical distribution, with due regard to the technological significance of each State Party's autonomous systems development programs and artificial intelligence research capacity, the volume of its defense-related autonomous systems acquisition, and the size of its armed forces. Members shall serve renewable terms of two years and shall not simultaneously represent a State Party that is the subject of compliance proceedings under Article 12.

3. The Technical Secretariat assists the Conference of States Parties and the Executive Council in the performance of their functions, carries out the verification measures, behavioral testing administration, and confidence-building measure collection provided in this Convention, and reports to the Executive Council on an ongoing basis. The Director-General, appointed by the Conference of States Parties

for a renewable term of four years, is the chief administrative officer and legal representative of IAWCO.

4. The Scientific and Technical Advisory Board advises the Technical Secretariat and the Conference of States Parties on all scientific and technical questions arising under this Convention. It shall consist of not fewer than twenty-five independent experts serving in their personal capacity, selected for expertise in artificial intelligence and machine learning architecture, cybersecurity and offensive cyber capabilities, critical infrastructure engineering, and international humanitarian law. Experts shall be nominated by States Parties and appointed by the Conference of States Parties to serve renewable three-year terms. No State Party shall hold more than three seats on the Board at any one time.

5. The Conference of States Parties shall convene quinquennial Review Conferences to assess the operation and status of this Convention; to evaluate whether advances in autonomous systems technology or explainable artificial intelligence have materially altered the definitional scope of Article 1(1) and 1(2) or the humanitarian risk profile upon which the categorical prohibition in Article 3 is premised; and to consider, in accordance with Article 11(2), whether the verification-impossibility premise underlying that prohibition requires reassessment. Review Conferences shall also assess the adequacy of the behavioral testing protocols established under Article 6 in light of technological developments since the preceding Review Conference.

Negotiating Commentary, Article 15: Article 15(1)'s quorum requirement, a majority of all States Parties present and voting, responds to a governance failure documented in CWC Conference practice: that substantive decisions with Convention-wide implications have occasionally been reached by thin participation, eroding

their legitimacy. The requirement ensures that the Conference's determinations carry the political authority necessary for effective implementation.

Article 15(2)'s twenty-one-member Executive Council is modelled on CWC Article VIII(23)–(25). The disqualification clause, precluding a State Party subject to compliance proceedings from simultaneously serving as an Executive Council member, supplements the CWC model and responds to the evident conflict of interest that would arise if a State facing compliance scrutiny retained executive decision-making authority over that scrutiny.

Article 15(4)'s Scientific and Technical Advisory Board constitutes the most consequential institutional innovation in this Convention. The OPCW's Scientific Advisory Board has demonstrated over three decades that independent technical expertise embedded within a treaty institution can maintain the technical relevance of an arms control regime as technology evolves, without requiring formal treaty amendment for each adaptation. The cap of three seats per State Party prevents technological leaders from structurally dominating the Board. The quinquennial Review Conferences in Article 15(5) follow the model established by the NPT (Article VIII(3)) and the CWC (Article VIII(22)), providing the institutional mechanism through which the Convention can adapt to technological change without requiring formal amendment for every development.

CHAPTER VIII
FINAL PROVISIONS

Article 16, Relationship with Other International Agreements

Nothing in this Convention shall be interpreted as limiting or derogating from obligations assumed by States Parties under Additional Protocol I to the Geneva Conventions of 12 August 1949, the Charter of the United Nations, the Rome Statute of the International Criminal Court, the Convention on Certain Conventional Weapons and its Protocols, or any other applicable rule of conventional or customary international law. This Convention is supplementary to, and not in derogation of, the rules of customary international humanitarian law, including those governing the prohibition on causing unnecessary suffering, the principles of distinction and proportionality, and the precautionary obligations applicable to the conduct of hostilities. Where this Convention establishes obligations more specific or more stringent than those arising under applicable international law, the more stringent obligation shall govern.

Negotiating Commentary, Article 16: Article 16 establishes the Convention's structural position within the international legal order as an instrument of addition and specification rather than a lex specialis displacing the general law of armed conflict. The clause "supplementary to, and not in derogation of" follows the standard savings clause formulation employed in AP I Article 75, ensuring that neither this Convention's more specific obligations nor its silence on any given question can be invoked to reduce protections available under existing IHL. Article 16's function as an instrument of normative addition is analogous in structure to Article 75 of AP I, which preserves rights and obligations arising under other instruments of international humanitarian law and does not displace more favorable protections established thereunder. This Convention governs

weapons design and pre-deployment review; IHL governs the conduct of each individual operation. A State may comply with every procedural obligation in this Convention and still violate AP I's principle of proportionality if its autonomous cyber weapon deployment produces effects disproportionate to the concrete and direct military advantage anticipated, or the principle of distinction if it produces effects against persons or objects that have lost or never acquired military character.

The principle of humanity, as confirmed by the ICJ in Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 1996 I.C.J. Rep. 226, ¶ 87 (July 8), which affirms that the Martens Clause constitutes 'an effective means of addressing the rapid development of military technology,' ensures that emerging weapons incapable of being governed by existing treaty provisions remain subject to the principles of humanity and the dictates of public conscience as operative sources of legal constraint. That constraint may impose requirements beyond those established in either this Convention or AP I, and Article 16's structure ensures that no argument from the Convention's silence displaces obligations derived from that principle.

The relationship between this Convention and the positions stated in the Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations (2017) and any successor instrument produced through the Tallinn process warrants explicit clarification. The Convention does not supersede, qualify, or derogate from any normative position advanced in those instruments; it establishes binding treaty obligations in a space where the Tallinn Manual process has articulated customary law positions but has not generated treaty commitments. The Convention's weapons review and behavioral testing obligations complement the due diligence standard articulated in Tallinn Manual 2.0 Rule 6, which requires States to take all feasible measures to ensure that their territory is not used for operations that

violate other States' international law obligations, a standard that the preventive architecture of this Convention operationalizes at the pre-deployment stage, converting the Tallinn Manual's customary due diligence norm into the specific institutional procedures of Articles 4 and 6.

Article 17, Signature, Ratification, Accession

This Convention shall be open for signature at the United Nations Headquarters in New York from the date of its adoption by the General Assembly. It shall be subject to ratification, acceptance, or approval by signatory States and shall be open for accession by any State that has not signed the Convention. Instruments of ratification, acceptance, approval, or accession shall be deposited with the Secretary-General of the United Nations, who is hereby designated as the Depositary of this Convention. The Depositary shall inform all signatory and acceding States, and the Director-General of IAWCO following its establishment, of the date and content of each such deposit.

Article 18, Entry into Force

This Convention shall enter into force 180 days after the date on which the fiftieth instrument of ratification, acceptance, approval, or accession is deposited with the Depositary. For each State subsequently depositing an instrument of ratification, acceptance, approval, or accession, the Convention shall enter into force 30 days after the date of deposit of that State's instrument.

Negotiating Commentary, Article 17 and 18: The entry-into-force threshold of fifty States Parties is calibrated below the CWC threshold of sixty-five and above the Ottawa Convention threshold of forty, balancing the need for critical mass, particularly among technologically capable States without whose participation the Convention cannot achieve its verification and behavioral testing objectives,

against the risk that an excessively high threshold defers entry into force indefinitely. The 180-day gap between deposit of the fiftieth instrument and entry into force provides the interval necessary for the inaugural Conference of States Parties to convene, elect the Executive Council, appoint the Director-General, and establish the Technical Secretariat's basic operational infrastructure before substantive obligations take legal effect.

Article 19, Amendments

1. Any State Party may propose amendments to this Convention. Proposed amendments shall be submitted in writing to the Director-General of the Technical Secretariat, who shall circulate them to all States Parties not later than ninety days before the session of the Conference of States Parties at which they are to be considered.

2. Amendments to the substantive provisions of this Convention, including those establishing prohibitions, definitions, and institutional obligations, shall be adopted by the Conference of States Parties and shall enter into force upon deposit of instruments of acceptance by two-thirds of all States Parties.

3. Amendments to the Annexes, including the Verification Annex and the behavioral testing protocols established under Article 6, shall be adopted by the Conference of States Parties by a two-thirds majority of members present and voting and shall enter into force 180 days after their adoption unless, within that period, more than one-third of all States Parties notifies the Director-General in writing of its objection to the amendment.

Negotiating Commentary, Article 19: The differentiated amendment procedure, requiring two-thirds acceptance of all States Parties for substantive amendments while permitting adoption of technical annex amendments by Conference majority with a silent-assent entry-into-force mechanism, follows the model established in CWC

Article XV. The differentiation reflects a practical necessity confirmed by the CWC negotiating record: the behavioral testing protocols in Article 6 and the Verification Annex procedures must be capable of technical adaptation as autonomous systems technology evolves without requiring each adaptation to obtain the acceptance of two-thirds of all States Parties, a threshold that would render technical updating functionally impossible at the pace of development that characterizes AI systems.

The substantive amendment procedure's two-thirds acceptance requirement ensures that changes to the Convention's core prohibitions, particularly any relaxation of the Article 3 categorical prohibition under Article 11(2)'s XAI reassessment mechanism, require a political consensus of genuine breadth and that no amendment removing or qualifying the central prohibition can be achieved without the support of a substantial majority of the treaty community. The ninety-day advance circulation requirement in paragraph 1 ensures that States Parties have adequate time to conduct national legal and technical review of proposed amendments before the Conference session at which they are to be considered, a minimum procedural guarantee whose absence from several arms control treaty instruments has generated legitimate complaints about insufficient deliberative opportunity. The silent-assent mechanism in paragraph 3 for Annex amendments, entering into force unless a majority objects, follows the precedent established in CWC Article XV(4) for Annexes to that Convention and places the procedural burden appropriately: those who object to a technical update must register that objection actively, rather than passive non-engagement defeating necessary technical adaptation.

Article 20, Withdrawal

Each State Party shall, in exercising its national sovereignty, have the right to withdraw from this Convention if extraordinary events related to the subject matter of this Convention have jeopardized its supreme national interests. It shall give notice of such withdrawal ninety days in advance to all other States Parties, the Depositary, and the United Nations Security Council. The notice shall include a statement of the extraordinary events the State Party regards as having jeopardized its supreme national interests. Withdrawal shall not affect any obligation incurred or arising prior to the effective date of withdrawal, including participation in compliance proceedings initiated before that date and the destruction obligations applicable to Tier 3 autonomous cyber weapons in the withdrawing State's possession at the date of notification or at any date during the ninety-day notice period, including weapons transferred to third parties during that period.

Negotiating Commentary, Article 20: The withdrawal provision follows the formulations of CWC Article XVI and NPT Article X(1) with deliberate textual fidelity, reflecting the standard structure that has evolved across the arms control treaty corpus and whose language carries recognized interpretive authority. The "supreme national interests" standard is drawn directly from NPT Article X(1) and has been applied in State practice as requiring a genuine and material threat to core national security interests, not a mere political disagreement with the Convention's obligations or dissatisfaction with the pace of other States' compliance. The requirement of notification to the United Nations Security Council, drawn from NPT Article X(1), subjects the withdrawal decision to the scrutiny of the organ charged under UN Charter Chapter VI with the peaceful settlement of disputes and under Chapter VII with responses to threats to international peace and security. The Security Council possesses

no veto over a State Party's decision to withdraw from this Convention, withdrawal is a sovereign right under VCLT Article 56 and, here, an express treaty provision, but the notification requirement ensures that the international community receives timely warning and that the Security Council may consider whether the asserted grounds for withdrawal raise issues of international peace and security requiring a response under the Charter.

The clause preserving obligations incurred prior to withdrawal is expanded relative to the CWC model to address an evasion strategy specifically adapted to the autonomous cyber weapons context. A State Party facing a non-compliance determination under Article 12, with attendant suspension of cooperative benefits and potential Security Council referral, might seek to withdraw from the Convention precisely to escape the compliance proceeding and its consequences. The explicit extension of the preservation clause to 'destruction obligations applicable to Tier 3 autonomous cyber weapons in the withdrawing State's possession at the date of notification' forecloses the parallel evasion strategy of withdrawing while retaining prohibited weapons stockpiles: the obligation to destroy declared Tier 3 autonomous cyber weapons survives the Convention's withdrawal as an obligation incurred during membership, binding on the withdrawing State as a matter of treaty law until discharged.

Article 21, Depositary Functions

The Secretary-General of the United Nations shall perform the depositary functions provided for in this Convention in accordance with Articles 76 and 77 of the Vienna Convention on the Law of Treaties of 23 May 1969. The Depositary shall promptly inform all signatory and acceding States of the date of each signature, the date of deposit of each instrument of ratification, acceptance, approval, or accession, the date of entry into force of this Convention and of

any amendment thereto, and the date and content of any notification of withdrawal.

Negotiating Commentary, Article 21: The designation of the UN Secretary-General as Depositary follows the model of the CWC (Article XXII), the Ottawa Convention (Article 20), and the NPT (Article IX(2)), all of which designate the Secretary-General in this capacity. The reference to VCLT Article 76 and 77, which establish the functions and duties of depositaries of multilateral treaties, ensures that the Depositary's obligations are defined by reference to the most authoritative codification of depositary practice in international treaty law, avoiding the need to enumerate those functions specifically in the Convention itself. The enumeration in the second sentence of the principal notification obligations, signature dates, ratification deposits, entry into force, and withdrawal notices, ensures that States Parties receive timely and complete information about the Convention's membership status, which is essential both for States deciding whether to accede and for IAWCO in determining the composition of the Conference of States Parties and the Executive Council.

Article 22, Authentic Texts

This Convention, of which the Arabic, Chinese, English, French, Russian, and Spanish texts are equally authentic, shall be deposited with the Secretary-General of the United Nations.

Negotiating Commentary, Article 22: The six authentic language texts correspond to the six official languages of the United Nations, following the standard formulation employed in CWC Article XXIII, Ottawa Convention Article 22, and Rome Statute Article 128. The equal authenticity of all six texts is a rule of treaty interpretation under VCLT Article 33(1): where a comparison of the authentic texts discloses a difference of meaning that Article 31 and

Article 32 do not resolve, the meaning that best reconciles the texts, having regard to the object and purpose of the Convention, shall be adopted. For a Convention whose core prohibitions rest on precisely defined technical concepts, network propagation capacity, temporal persistence, irreversible effects, attribution resistance, the risk of divergence between language versions is material, and the Scientific and Technical Advisory Board established under Article 15(4) shall be available to advise the Conference of States Parties on terminological questions arising from such divergence.

Concluding Doctrinal Integration Note

Annex I constitutes the dissertation's primary *de lege ferenda* instrument. It demonstrates that the normative conclusions advanced across Chapters 2 through 6 are susceptible of concrete implementation through the existing multilateral treaty-making mechanisms of the United Nations system, drawing on verified institutional precedents rather than theoretical propositions alone. The Convention's architecture, absolute categorical prohibition sealed against *jus contra bellum* circumvention, mandatory pre-deployment weapons review extended to AI-assisted semi-autonomous systems, behavioral testing aligned to the three operational characteristics that define the prohibited category, together with the attribution-resistance factor addressed as an aggravating circumstance under Article 3(6), institutional verification through a three-tier inspection regime, a self-correcting quinquennial review mechanism anchored to the State of explainable AI technology, and individual criminal responsibility reoriented to the deployment authorization stage, operationalizes the dissertation's central thesis that the accountability gap generated by the temporal and epistemic disaggregation of autonomous targeting decisions cannot be remedied through reactive application of

existing doctrine, but demands preventive governance embedded at the development and acquisition stage.

This Convention provides the institutional framework within which it can be operationalized.

- End of Annex I -

STATUTORY DECLARATION

(Eidesstattliche Versicherung)

Ich versichere an Eides statt, dass ich die eingereichte Dissertation selbständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Ort, Datum: Unterschrift:

Berlin, den 26. Juli 2026