Reconstructing Responsibility in Algorithmic Warfare: Gaza, Libya/Kargu-2, Air-Defense Systems, and the Control Envelope under International Law

إعادة بناء المسؤولية في الحرب الخوارزمية: غزة، ليبيا/كارجو-2، أنظمة الدفاع الجوي، ونطاق السيطرة التشغيلية في القانون الدولي

MAHA Zeggagh *
LSRMPUDI Research Lab
University of Batna 1
maha.zeggagh@univ-batna.dz

ADEL Zeggagh
LSRMPUDI Research Lab
University of Batna 1
adel.zeggagh@univ-batna.dz

Date of Submission 02-04-2026 Date of Final Acceptance 28-07-2026 Date of Publication September 2026

Abstract:

This article examines responsibility under international law for autonomous and AI-enabled military systems as algorithms play a growing role in identifying, ranking, and prioritizing targets. It asks whether existing legal rules can still support attribution, knowledge, foreseeability, traceability, and meaningful human control when human involvement is reduced, divided among several actors, or limited to approving system outputs. The discussion draws on international humanitarian law, international criminal responsibility, and state responsibility, and applies these frameworks to three cases: AI-assisted targeting in Gaza, autonomous engagement in Libya/Kargu-2, and automated air-defense systems. The cases show that accountability problems often stem from weak practical conditions for applying existing law. The article therefore proposes the control envelope as a standard for defining lawful operational limits and argues that systems which cannot be kept within those limits should be strictly restricted or prohibited in relevant contexts.

Keywords: Algorithmic Warfare; International Responsibility; Autonomous Weapon Systems; Meaningful Human Control; Control Envelope.

ملخص:

يتناول هذا المقال المسؤولية في القانون الدولي عن الأنظمة العسكرية الذاتية والمدعومة بالذكاء الاصطناعي، في ظل تزايد اعتماد العمليات العسكرية على الخوارزميات في تصنيف الأهداف وترتيبها وتسريع القرار القتالي. ويبحث في مدى قدرة القواعد القائمة على ضمان الإسناد والعلم وقابلية التوقع والتتبع والتحكم البشري الهادف، عندما يتقلص الدور البشري أو يتوزع بين جهات متعددة أو ينحصر في الموافقة على مخرجات النظام. وينطلق التحليل من قواعد القانون الدولي الإنساني والمسؤولية الجنائية الدولية ومسؤولية الدولة، ثم يطبقها على ثلاث حالات: الاستهداف المدعوم بالذكاء الاصطناعي في غزة، والاشتباك الذاتي في ليبيا/كارجو-2، وأنظمة الدفاع الجوي المؤتمتة. وتبين هذه الحالات أن صعوبات المساءلة ترتبط غالبا بضعف الشروط العملية اللازمة لتطبيق القواعد القانونية. لذلك يقترح المقال نطاق السيطرة التشغيلية معيارا يحدد حدود الاستخدام المشروع، ويوصي بتقييد أو حظر الأنظمة التي يتعذر ضبطها ضمن هذا النطاق.

الكلمات المفتاحية: الحرب الخوارزمية؛ المسؤولية الدولية؛ أنظمة الأسلحة الذاتية؛ التحكم البشري الهادف؛ نطاق السيطرة التشغيلية.

Corresponding Author: MAHA Zeggagh

Introduction

Autonomous and AI-enabled military systems are often approached through a single question: who is responsible when a machine participates in the use of force? This question captures only part of the responsibility problem on the contemporary battlefield. Alongside weapons capable of selecting and engaging targets after activation, states increasingly use AI-assisted targeting systems that classify and rank candidates for human approval, loitering munitions that continue the targeting process after launch, and automated defensive systems that operate at speeds difficult for human judgment to follow. Human decision-makers remain involved in most of these settings, but their role changes considerably. Speed may restrict deliberation, responsibility may be distributed across several institutions, and decisive choices may already be embedded in target profiles, technical thresholds, system settings, or operational doctrine before final approval is given.[1]

The literature has identified the risk of a responsibility gap with considerable force. Crootof showed that autonomous weapons strain existing legal categories when unlawful harm cannot easily be connected to individual criminal fault or command responsibility.[2] De Vries and Kwik refined the debate by arguing that difficulties of attribution do not always amount to a legal vacuum, and that some gaps are evidentiary, institutional, or manufactured through poor review and avoidable ignorance.[3] Others have turned to meaningful human control, state responsibility, weapons review, and the political character of technical design.[4] Although these studies clarify important parts of the problem, they have not yet been brought together into a single operational account of responsibility across different forms of algorithmic mediation.

The significance of this study lies in addressing the growing mismatch between the formal attribution of legal responsibility and the operational conditions under which algorithmic systems shape decisions involving lethal force. The difficulty is no longer confined to determining whether a human participated in the decision. It also concerns whether that participation occurred under conditions that allowed the human actor to understand the system's output, assess the relevant risks, challenge the recommendation, intervene where necessary, and remain accountable for the resulting decision.

This article asks how international law can reconstruct responsibility for autonomous and AI-enabled military systems when algorithmic mediation compresses, redistributes, or formalizes human agency rather than removing it. The article proceeds from the hypothesis that the principal responsibility gap arises when institutions fail to preserve foreseeability, traceability, reviewability, and meaningful human control throughout the system's operational life cycle, even though the relevant rules of international law remain applicable. International humanitarian law, individual criminal responsibility, command responsibility, state responsibility, and weapons review obligations continue to govern these systems. Their application becomes difficult when the system's function, target profile, data inputs, error rates, review process, human role, and records are vague or inaccessible, because the conditions needed to establish knowledge, control, causation, and attribution are then weakened.

The article combines doctrinal legal analysis with a structured comparative inquiry into three distinct forms of algorithmic mediation. AI-assisted targeting in Gaza examines how reported systems such as Lavender and The Gospel (Habsora) may affect the distinction between machine-generated classifications and legal determinations under international humanitarian law, particularly when human review is compressed by speed and scale. Libya/Kargu-2 examines autonomous engagement after activation and the extent to which responsibility can be reconstructed through mission parameters, target definitions, operational constraints, environmental suitability, and

traceability. Automated air-defense systems examine the limits of formal human-in-the-loop control where speed, interface design, uncertainty, and automation bias may reduce independent judgment.[5]

The cases differ in their technology and operational setting and are not used to reach definitive factual conclusions about particular attacks. They are examined through common analytical criteria: the function assigned to the system, the definition of the target category, the suitability of the operational environment, the quality of human review, the visibility of uncertainty and error, the preservation of records, and the attribution of knowledge and control across the institutional chain.

The comparison leads to the article's central proposal: the control envelope, understood as the documented operational, technical, and legal boundaries within which an autonomous or AI-enabled military system may be used with a reasonable degree of foreseeability, traceability, reviewability, and human accountability. The envelope requires the authority responsible to define and document the system's function, permissible targets, operational environment, human role, known limitations, review procedures, and record-keeping duties. It does not create a new legal regime. It provides a framework through which existing rules can be applied to the institutional and technical chain that precedes and follows the use of force. It also allows investigators, commanders, legal reviewers, and courts to determine who authorized the system, what risks were known, what restrictions governed its operation, and whether human judgment remained capable of carrying legal responsibility.[6]

The article proceeds in five sections. Section I develops the conceptual and legal framework by linking responsibility gaps to the practical conditions of legal accountability. Sections II, III, and IV examine AI-assisted targeting in Gaza, autonomous engagement in Libya/Kargu-2, and automated air-defense systems. Section V reconstructs individual, state, and institutional responsibility and develops the control envelope as an operational standard. The conclusion presents the principal findings and recommendations.

Section I: From Responsibility Gaps to the Conditions of Legal Accountability

This section establishes the conceptual and legal framework for examining responsibility in autonomous and AI-enabled military operations. It begins by distinguishing the absence of applicable legal rules from failures in the practical conditions needed to assign responsibility. It then examines meaningful human control, weapons review, state responsibility, and the functions delegated to algorithmic systems before considering how the principles of distinction, proportionality, and precautions can be applied in algorithmically mediated operations

A) From Responsibility Gaps to the Practical Conditions of Accountability

Robert Sparrow's early critique remains the reference point for the debate. Sparrow argued that deploying autonomous weapons may leave no human agent who can be held morally responsible in a satisfactory way for wrongful harm.[7] Rebecca Crootof translated the concern into legal terms, showing that unlawful harm caused by autonomous weapons may fit awkwardly into conventional models of individual fault or command responsibility.[8]

Later scholarship resisted the inference that autonomy automatically produces a legal vacuum. Barry de Vries distinguishes apparent from real responsibility gaps, tracing many difficulties to evidentiary opacity or causal complexity rather than to the absence of applicable law.[9] Jonathan Kwik argues that the gap should not be treated as a single unavoidable consequence of autonomy[10], and his later work on ignorance goes further, showing that accountability gaps may be produced or aggravated by poor institutional choices, inadequate testing, and cultivated ignorance.[11]

Taken together, their work redirects the debate toward the conditions required for responsibility and whether those conditions were maintained in practice.

B) Meaningful Human Control beyond Formal Human Presence

Another strand of literature examines the problem through the idea of human control. The ICRC has insisted that human control remain central to the development and use of autonomous weapon systems, particularly in the critical functions of selecting and engaging targets.[12] Amoroso and Tamburrini structure that requirement by distinguishing preventive, epistemic, and legal dimensions of meaningful human control, so that control is not exhausted by human presence at the moment of attack but also demands prior constraints, knowledge of the system, legal supervision, and a link between the system's operation and accountable human judgment.[13]

Dan Saxon grounds the critique in human dignity, arguing that delegating life-and-death judgments to autonomous systems erodes the human responsibility that gives international law its moral significance. His most useful move is to identify speed, rather than autonomy alone, as the central threat to human judgment: a human formally in or on the loop may have no meaningful time for understanding, deliberation, or intervention.[14] Nathan Wood challenges strong versions of the control requirement from the opposite direction, arguing that control need not be contemporaneous with each attack, since a system may extend human will where it operates within ex ante operational constraints, targeting parameters, and a defined policy framework. Wood's distinction between systems that extend human will and those that offload it nonetheless supports the same basic concern, because it makes responsibility depend on whether human judgment is genuinely structured into the system or merely displaced by it.[15]

C) Weapons Review and State Responsibility as Conditions of Foreseeability

Afonso Seixas-Nunes insists that AWS are weapon systems rather than new battlefield agents, which forecloses the misleading idea that machines might become legal subjects. He nonetheless shows how accountability becomes difficult when harm arises from mission algorithms, software errors, environmental interaction, or the improper encoding of IHL and rules of engagement.[16] His treatment of dolus eventualis and recklessness supplies a doctrinal route for cases in which commanders, programmers, or technical personnel knowingly accept a substantial risk of an IHL violation[17], while his warning against manufacturing scapegoats where no individual possessed the relevant knowledge or control opens the way to a stronger role for state responsibility.[18]

Lutiana Barbosa develops that role, asking whether existing attribution rules, especially as reflected in ARSIWA, are adequate for harms involving autonomous weapon systems.[19] Traditional attribution presupposes a model of human conduct, whereas AWS-related harms emerge from interactions among software, data, sensors, operators, military institutions, and private contractors.[20] Barbosa therefore emphasizes due diligence, weapons review, human-machine interaction, testing, auditability, and the management of private suppliers.[21] Her analysis places the emphasis on whether the state established reasonable conditions of foreseeability before deployment, alongside the later question of individual culpability.

Definitional choices carry comparable weight. Taddeo and Blanchard show that official definitions of AWS diverge sharply, some emphasizing autonomy, others adaptive capability, human control, or purpose of use.[22] *Berenike* Prem adds a caution against the imaginary of "control-by-design": target profiles, error rates, validation procedures, and testing environments encode political, military, and legal assumptions into technical criteria.[23] Technical design can support

responsibility when it is accompanied by institutional review, documentation, opportunities to challenge outputs, and procedures for modifying or withdrawing the system when its assumptions prove unreliable; otherwise it legitimizes autonomy while offering fragile assurance. Work on decision-support systems completes the picture, since AI tools that generate target lists, assign risk scores, and estimate collateral damage narrow the practical space for legal judgment under speed, opacity, workload, and institutional pressure even where the human remains formally present.[24]

D) Functional Delegation beyond Formal System Labels

A sound legal analysis should begin with the functions performed by the system rather than with its formal label. Responsibility arises from the specific function delegated to a system, the operational environment of its use, the degree of human judgment retained, and the extent to which behavior remains foreseeable and traceable.

The most common functional marker is the delegation of "critical functions": the selection and engagement of targets. Neil Davison and the ICRC use this language to separate systems that assist human operators from those performing functions directly linked to the application of force.[25] The formulation still requires unpacking. Selection may involve detection, identification, classification, ranking, prioritization, or the matching of a sensor input to a target profile; engagement may involve tracking, firing, loitering, or triggering an effect once a condition is satisfied. The legal inquiry must identify the part of the targeting process assigned to the system, the limits imposed on that function, and the degree of human supervision that remains possible.

Seixas-Nunes clarifies this by organizing AWS around mission algorithms comprising situation assessment, selection, and situation management.[26] A harmful strike may be shaped long before the final moment of attack, by how the system assesses the environment, interprets sensor data, applies a target profile, or adapts to battlefield conditions. Reconstructing responsibility accordingly requires asking who defined the mission, encoded the constraints, approved the data environment, tested the system, and set the conditions under which it could operate.

AI-enabled targeting systems must therefore be included in the analysis, while remaining conceptually distinct from autonomous weapon systems. A decision-support system may never engage a target and still reshape the legal decision by producing lists, scores, estimates, or recommendations that operators are institutionally pressured to accept. Wood's distinction captures the difference: a system extends human will when it operates within parameters deliberately set by competent human agents, and offloads it when the human becomes dependent on outputs he cannot meaningfully understand, contest, or revise.[27]

The familiar categories of human-in-the-loop, on-the-loop, and out-of-the-loop remain useful descriptive tools, although they do not by themselves establish whether human involvement is legally meaningful.[28] *Saxon's* caution about speed applies throughout: where operational tempo deprives humans of deliberative space, the legal significance of their formal presence becomes doubtful.[29]

E) Translating IHL Obligations into Algorithmic Operations

IHL remains the governing framework, and autonomy displaces none of it. Distinction, proportionality, and precautions in attack bind any state deploying AWS or AI-enabled targeting tools. The main difficulty lies in translating these legal obligations into the technical and operational processes through which algorithmic systems function, since IHL rules are structured around contextual legal judgment while algorithmic systems operate through data, models, classifications,

thresholds, and parameters. A recurring risk arises when a probabilistic classification is treated as if it were itself a legal determination.

This problem appears most clearly in relation to the principle of distinction. A machine-learning system may classify a person, vehicle, building, or signal pattern as matching a target profile — a classification that can be statistically useful without amounting to the legal conclusion that the object is a military objective or the person targetable. Prem explains why: target profiles translate military assumptions, legal judgments, intelligence categories, and operational priorities into machine-readable criteria.[30] Criteria that are broad, weakly validated, or detached from context yield outputs that appear technically reliable while remaining legally underdetermined. Legal evaluation must consider both the model's technical performance and whether its classifications can support the contextual judgment required by international humanitarian law.

The principle of proportionality presents a similar difficulty because its application cannot be reduced to a numerical calculation. Its vocabulary invites quantification — expected military advantage weighed against expected incidental harm — yet the test is no numerical ratio. It involves qualitative judgment under uncertainty, made by a commander on the basis of information reasonably available at the time. AI systems can estimate patterns, identify objects, and model effects; they cannot determine whether expected harm is excessive in relation to anticipated military advantage. The more an institution relies on automated estimates or thresholds, the more it matters who defined those thresholds, how uncertainty was represented, and whether reviewers had authority and time to reject the recommendation.

The obligation to take precautions in attack provides the clearest connection with the framework developed in this article. The obligation requires verification of targets, choice of means and methods that minimize civilian harm, and cancellation or suspension when the target proves unlawful or the expected harm excessive. Satisfying it in algorithmic warfare takes more than placing a human somewhere in the process; it requires a decision architecture in which verification, review, escalation, suspension, and post-strike assessment are materially possible. Where a system compresses the kill chain so far that no meaningful review can occur, the question becomes whether the state or command structure designed a process incompatible with the practical exercise of precaution.

Article 36 weapons review is central here, but not if understood as a one-time certification of a technical object. Reviewing a weapon, means, or method of warfare requires asking whether its employment would be prohibited in some or all circumstances — which, for AWS and AI-enabled systems, means examining the intended mission, training data, target profiles, human-machine interface, environmental assumptions, update mechanisms, cybersecurity vulnerabilities, testing conditions, and operational doctrine rather than the platform in isolation. Review and human-machine interaction form part of the legal infrastructure through which a state builds, or fails to build, foreseeability.[31] A review that remains limited to formal approval may recognize the applicable law without clarifying the conditions under which the system will actually be used.[32]

The case studies that follow are guided by three conclusions. Autonomy is legally meaningful only when connected to specific delegated functions. Human involvement is legally meaningful only when it preserves the capacity for informed judgment, intervention, and accountability. Responsibility is legally meaningful only when the state and its military institutions have built the conditions of foreseeability and traceability before deployment.

Section II: Gaza: Algorithmic Classification and the Compression of Legal Review

Gaza is relevant to this debate because it illustrates a different form of algorithmic involvement from the classical autonomous weapon. It shows how responsibility can be strained while a human remains formally present in the targeting chain. Algorithmic systems classify persons and objects, generate target lists, estimate risk, accelerate operational tempo, and leave reviewers with too little time, information, or institutional freedom to exercise legal judgment. The case therefore shows how responsibility may be weakened when human judgment is compressed and reduced to formal approval.

Open-source investigations by +972 Magazine, Local Call, and The Guardian reported that the Israeli military used AI-assisted systems during the Gaza campaign. Lavender was described as identifying suspected members of armed groups, while The Gospel (Habsora) was described as generating or recommending targets, particularly buildings and other structures. The reports alleged that large numbers of potential targets were generated and that some human reviews were extremely brief.[33] *This* article does not adjudicate those factual disputes. It uses the reported architecture as a legally relevant configuration: algorithmic classification followed by human approval under conditions of speed and scale.

This configuration calls into question the common assumption that responsibility is preserved whenever a human remains in the loop. Where the reviewer sits at the end of a process in which the system has already selected, ranked, and framed the target, the reviewing role becomes confirmatory rather than deliberative, and the legal question becomes whether that human had enough time, information, training, authority, and institutional independence to question the output. An AI-assisted targeting system may be no weapon at all in the narrow sense, yet it mediates the legal decision by transforming raw intelligence into targetability signals that reach the reviewer as a probability score, a category, a recommendation, a ranked list, or a collateral-damage estimate. The analysis must preserve a clear distinction between algorithmic classification and legal determination: a model may classify an individual or structure as matching a target profile without establishing that the person is targetable under IHL or that the object is a military objective.[34]

Prem's critique of target profiles explains the risk. Profiles translate assumptions about behavior, identity, affiliation, location, communication patterns, and military relevance into machine-readable criteria.[35] Where those criteria are broad, weakly validated, or poorly matched to the operational environment, outputs look precise while remaining legally fragile. Dense urban environments sharpen the difficulty, since civilian life, military activity, communications, infrastructure, and movement patterns overlap. Distinction in such settings cannot be reduced to pattern recognition; it demands contextual judgment about status, conduct, object, use, contribution to military action, and the circumstances prevailing at the time of attack.

The first legal difficulty concerns the principle of distinction. Where an algorithmic system generates a list of suspected persons or structures, the human decision-maker must still ask whether the available information supports targetability under IHL, weighing the evidentiary threshold, the type and recency of the intelligence, the system's error rate, the possibility of misclassification, and the availability of contrary information. A target list is at most an input into legal judgment.

A second difficulty concerns proportionality. AI-assisted systems can estimate collateral harm, model building occupancy, and supply numerical risk indicators, all of which may be useful. Yet the commander must still assess whether expected incidental civilian harm would be excessive in relation to the concrete and direct military advantage anticipated — an assessment whose

qualitative and contextual elements no score can fully capture. Where an institution converts numerical thresholds into routine authorization rules, proportionality risks becoming administratively standardized rather than legally judged, and defective thresholds may normalize patterns of harm even as they improve consistency.

The third issue relates to the practical exercise of precautions in attack. When reviewers receive a large volume of machine-generated files, where the interface highlights confidence scores without explaining uncertainty, where institutional tempo rewards approval over challenge, or where the reviewer has only a brief window to act, the precautionary function is hollowed out. The presence of a human signature in the chain does not answer the question; the capacity of the review architecture to support precautionary judgment does.

Rubber-stamp review becomes legally significant when approval is too limited to support independent judgment. Rubber-stamping does not mean the human is absent. It means the act of approval becomes too thin to carry the normative weight assigned to it. A reviewer who confirms a superficial attribute, accepts a system-generated category, or processes files at a pace incompatible with independent assessment preserves the appearance of human control while losing its substance. Saxon's argument about speed applies squarely: the compression of decision time erodes deliberative capacity even where formal procedures remain intact.[36] The automation-bias literature converges, since operators trained or pressured to trust a system, under workload and time pressure, become less able to reject erroneous outputs.[37]

Responsibility must therefore be examined across the wider institutional process rather than confined to the officer who gives final approval. Where that officer works within a system that has already defined target profiles, set thresholds, trained models, filtered data, generated rankings, estimated collateral harm, and imposed an operational tempo, responsibility is distributed across a wider institutional architecture: intelligence officers who build the profiles, commanders who authorize the workflow, technical personnel who validate the model, legal advisers who approve the methodology, procurement and data authorities who supply the system, senior decision-makers who set the pace and the permissible risk posture. Not all of these actors bear criminal responsibility, but all belong in the chain of institutional design and authorization that responsibility analysis must track.

Seixas-Nunes's mission-algorithm framework guides that tracking. If algorithmic military action comprises situation assessment, selection, and situation management, each stage must be examined.[38] In Gaza-type targeting the decisive legal failure may occur at assessment or selection rather than at weapon release: an unreliable classification at the assessment stage, or a weak correlation converted into a recommendation at the selection stage, may leave the final approval too late to repair the defect. The relevant inquiry is whether the system allowed meaningful verification before its recommendation was converted into operational action.

From the standpoint of state responsibility, the case points toward a duty to structure AI-assisted targeting so as to preserve foreseeability and reviewability. A state deploying such tools must be able to show that it reviewed the system, understood its limitations, validated it against realistic operational scenarios, defined permissible uses, trained users, monitored outputs, and retained auditable records.[39] A state unable to reconstruct how a target was generated, what data supported it, what uncertainty attached to it, who reviewed it, and why the recommendation was accepted has already weakened the conditions of accountability before the legality of any particular strike is assessed. Article 36 review must therefore extend beyond the weapon that delivers kinetic force, since a system that never fires may still shape the method by which targets are selected.

Review must consider data sources, model performance, known error rates, target-profile design, interface effects, human workload, escalation procedures, audit logs, and the capacity to suspend the system when outputs become unreliable.

These concerns do not make AI-assisted targeting inherently unlawful or human review necessarily ineffective. The Gaza case makes the more limited point that the legal value of review depends on the conditions in which it is carried out. Where algorithmic classification, institutional tempo, and formal approval combine, the responsibility problem migrates from the absence of a human to the quality of the human role, and the decisive question becomes whether human and institutional actors preserved the conditions under which distinction, proportionality, precautions, and accountability could be meaningfully exercised. Such systems demand a strict envelope: clear functional limits, legally defensible target profiles, validated data, meaningful review time, explainable uncertainty, audit trails, and an institutional culture that permits refusal rather than rewarding throughput. Without these safeguards, the law continues to apply, but its effective implementation becomes significantly more difficult.

Section III: Libya/Kargu-2: Foreseeability and Attribution after Autonomous Engagement

Libya/Kargu-2 raises the classical AWS problem that Gaza does not: what happens when a weapon system can search, select, and engage targets after activation without continuous connectivity to a human operator? Because the facts remain contested, the case is used here to identify the legal questions raised by autonomous engagement rather than to establish the first unlawful autonomous attack. What must be known before deploying such a system? How far must its mission be constrained? Who bears responsibility when the system's behavior is difficult to reconstruct? And should responsibility be sought in the last human operator or in the wider institutional decision to deploy?

The case entered the debate through the 2021 report of the UN Panel of Experts on Libya, which described the use of STM Kargu-2 loitering munitions and other systems during hostilities. The report stated that retreating forces were "hunted down and remotely engaged" by unmanned combat aerial vehicles or lethal autonomous weapons systems, and that the Kargu-2 could be programmed to attack targets without requiring data connectivity between operator and munition — a capacity it characterized as "fire, forget and find".[40] *Libya/Kargu-2* functions here as a stress test for foreseeability and attribution rather than as a proven fact pattern.

Kargu-2 does not fit neatly within the categories commonly used in legal analysis. It is neither an automated defensive system reacting within a fixed environment nor a science-fictional robot exercising independent moral agency, but a loitering munition with autonomous functions that may be configured to search for and engage targets within a mission envelope. Seixas-Nunes's insistence that AWS remain weapon systems rather than legal agents holds[41]: the system does not become a subject of law. But once a weapon can continue the targeting process after launch, responsibility analysis must move upstream — to the mission, the parameters, the target profile, the environment, the testing, the operator's role, and the decision to deploy in that context.

The first legal issue concerns foreseeability. Autonomy does not by itself make responsibility impossible. A state or commander may retain meaningful control by deploying the munition within a narrow, well-understood mission: a defined target type, a limited area, a limited duration, validated sensors, predictable behavior, abort conditions, adequate records. Wood's will-extending category covers precisely this, where human agents deliberately constrain operation through ex ante parameters that render outputs reasonably foreseeable.[42] The position changes when

the system is used in a fluid environment with broad target criteria, limited communications, uncertain sensor performance, and weak post-strike traceability. In such circumstances, the munition begins to look less like an extension of human will and more like a delegation of lethal judgment under conditions the human actor cannot meaningfully control.

Foreseeability, so understood, does not demand certainty. No weapon produces perfect predictability; conventional weapons malfunction, miss, and generate unintended effects. What matters legally is whether the relevant risks were reasonably knowable and whether they were managed through feasible precautions. Kwik resists the assumption that responsibility gaps follow inevitably from complexity, since many gaps are produced by how systems are reviewed, deployed, and documented.[43] Applied to Kargu-2, the question is not whether every movement could be predicted but whether the deploying authority confined the system to conditions in which its target selection and engagement behavior remained foreseeable enough for legal responsibility to attach.

A second issue concerns the attribution of the resulting conduct. Where a human operator directly selects and fires, attribution to human conduct is straightforward. Where a loitering munition continues searching after activation and engages on the basis of onboard processing, the chain lengthens: the relevant human act may be the launch, the mission programming, the choice of mode, the definition of the target profile, the approval of the operational area, the prior legal review, or the command decision to use the system in that environment. Barbosa shows why this matters, since ARSIWA's attribution framework is built around conduct by state organs or persons whose acts are attributable to the state, while AWS obscure the visible connection between human conduct and harmful outcome.[44] The appropriate legal response is to reconstruct the institutional decisions that enabled the system to operate, rather than treating the machine as an independent actor.

The case also demonstrates the limits of an approach that concentrates only on the final human operator. Where the system was launched in a mode permitting autonomous search and engagement, the decisive choices preceded the encounter with the target. Who authorized that mode? Who defined the target signature? Who determined the geographical box? Who assessed whether the system could distinguish lawful targets from civilians or civilian objects in that environment? Who conducted the weapons review? Who retained logs of what the system detected and why it engaged? The responsibility problem is documentary and institutional as much as causal: without records, parameters, and review trails, the legal system may know that a state used the system while remaining unable to reconstruct how the specific target became targetable.

Article 36 review is directly tested. Reviewing a loitering munition cannot stop at the platform's physical characteristics or the manufacturer's description; it must examine modes of operation, degrees of autonomy, target-recognition logic, sensor limitations, loss-of-link behavior, human override capacity, mission constraints, and the intended environment.[45] Where a state approves a system for narrow conditions and commanders deploy it under broader or more chaotic ones, the legal problem shifts from the abstract legality of the weapon to unlawful or negligent use. Where the review itself fails to identify the risks of autonomous mode, responsibility attaches at the institutional level.

The distinction between the legality of a weapon and the legality of its use sharpens accordingly. A loitering munition with autonomous functions is not unlawful per se; its lawfulness depends on whether it can be used consistently with distinction, proportionality, and precautions in the circumstances of use. Taddeo and Blanchard's emphasis on purpose of use and human control does analytical work.[46] A system configured to strike a clearly defined radar emitter in an isolated area raises different questions from one configured to search for persons or vehicles in a populated,

confused, or retreating environment. As the target category broadens and the environment grows more dynamic, the required control envelope becomes more demanding.

For individual criminal responsibility the case shows both possibility and difficulty. A commander who knowingly deploys an autonomous munition where it cannot distinguish lawful targets may incur responsibility if the material and mental elements are satisfied, and a technical officer or programmer may become relevant where he knowingly encoded or ignored a substantial risk that the system would attack protected persons or objects; dolus eventualis and recklessness supply the doctrinal route.[47] The same case warns against overextension, since forcing criminal responsibility onto the nearest operator, where no individual held the relevant knowledge, authority, or capacity, produces a scapegoat rather than accountability.

State responsibility occupies a central place in this analysis. The state or armed authority that procures, deploys, authorizes, or supplies an AWS-like system controls the institutional conditions of use even where the final engagement is algorithmically mediated, which is why the analysis must focus on due diligence, review, testing, control of contractors, and the decision to expose civilians or combatants to a system whose behavior is only partially foreseeable.[48] A state that benefits from the operational speed and deniability of autonomous functions should not be permitted to treat the resulting opacity as an excuse. Nor may it outsource accountability to vendors: where a system relies on proprietary software, undisclosed training data, or manufacturer-controlled updates, the state must still ensure that it can review, test, audit, and constrain the system before use, since private technical opacity otherwise becomes public legal opacity.

The same problem also has an important evidentiary dimension. A conventional strike can usually be reconstructed from orders, coordinates, munitions used, communications, and battle damage assessment. An autonomous engagement requires more: sensor inputs, classification outputs, confidence scores, mode settings, communication status, target-profile matches, system logs, abort opportunities. Where those records do not exist or are not preserved, the responsibility gap is in part a traceability gap — which does not make responsibility disappear but shows that the state failed to build the evidentiary architecture through which responsibility could be assessed. That failure should itself carry legal weight.

The operational environment remains central to the assessment. A system performing acceptably on a controlled test range may behave differently amid dust, smoke, damaged vehicles, irregular movement, mixed civilian and military traffic, spoofing, camouflage, or communication loss, and technical assurances built in design and testing grow fragile in adversarial and dynamic settings.[49] The relevant boundaries must therefore be dynamic rather than static, specifying not only where the system may be used but when changing conditions require suspension, human intervention, or withdrawal. This also separates two kinds of uncertainty. Uncertainty inherent in warfare does not automatically defeat legality and may be tolerable where managed through feasible precautions. Uncertainty created by deploying a system in conditions for which it was never adequately reviewed, tested, or constrained is closer to institutional negligence or reckless deployment. The line between them runs through what the deploying authority could reasonably anticipate and manage, which makes foreseeability a legal and institutional judgment rather than a purely technical metric.

Libya/Kargu-2 therefore supports the article's central diagnosis. Attribution weakens because action is distributed across human and algorithmic stages; foreseeability weakens because behavior depends on data, environment, and mode; criminal responsibility weakens because knowledge and control may not coincide in any single individual; and state responsibility becomes

crucial because the state controls the architecture of deployment. Legal responsibility depends on whether human and institutional actors defined, constrained, and documented the conditions under which the system could engage targets in a foreseeable, traceable, reviewable, and attributable manner.

Section IV: Air-Defense Systems: Automation Bias and the Limits of Formal Human Control

Air-defense systems offer a different perspective on the relationship between automation and responsibility. They involve neither AI-assisted target generation in a dense intelligence environment nor a munition searching for targets after launch. Their significance lies in the coexistence of formal human presence with a practical loss of judgment under speed, uncertainty, interface pressure, and automation bias. A human may sit in or on the loop while the system's tempo and presentation of information make independent legal judgment nearly impossible.

Most such systems are better described as automated or highly automated, operating according to pre-set rules, sensor inputs, and engagement logics rather than as contemporary AWS. They are useful for comparison because long-standing military systems show that responsibility may be weakened even before full autonomy is reached, especially when speed and machine-generated classification overwhelm human judgment.[50]

Their military rationale stems from the speed of the threats they are designed to counter. Incoming missiles, rockets, aircraft, and drones may move too fast for ordinary deliberation, and systems such as Phalanx and other close-in weapon systems were developed precisely because human reaction time is inadequate in last-resort defensive contexts.[51] Air defense therefore presents a particularly difficult problem for meaningful human control. Interpreting meaningful human control as direct deliberation over each individual engagement would render some defensive systems practically impossible; reducing it to prior activation and passive monitoring would empty the requirement. The task is to determine when speed justifies ex ante control and when it instead becomes a reason to prohibit or restrict particular modes of operation.

Wood's distinction between the human in the "fire loop" and the human in the "policy loop" is especially useful.[52] In a narrow defensive setting, control may be exercised in advance through the definition of protected assets, threat categories, geographic sectors, activation modes, rules of engagement, and shutoff or override conditions, allowing the system to extend human will without human deliberation over each microsecond engagement.[53] This form of prior control becomes insufficient when threat classification is uncertain, civilian or friendly aircraft share the same airspace, or operators cannot understand the basis of a recommendation.

Several historical incidents demonstrate the practical consequences of this problem, and they reveal how humans may remain formally responsible while relying on automated classifications they cannot adequately verify in the time available.[54]

Distinction appears here as a problem of identification: is the object an incoming hostile missile, a military aircraft, a civilian aircraft, a friendly aircraft, or an ambiguous radar track? A system may classify a track by speed, altitude, trajectory, transponder response, radar signature, or behavior — and the resulting label still falls short of a legal determination that the object may be attacked. As airspace grows more ambiguous, the verification requirement grows more demanding. Where the environment includes civilian traffic, friendly forces, or non-hostile aircraft, the control architecture must provide procedures for uncertainty, delay, escalation, and non-engagement, or the system will convert an uncertain sensor picture into an authoritative threat label.

Precautions must likewise be adapted to time pressure rather than abandoned to it. IHL demands no impossible certainty, but it does require feasible precautions to verify targets and minimize incidental harm — which in air defense may take the form of identification-friend-or-foe procedures, no-fire zones, civil aviation coordination, mode restrictions, operator training, human confirmation rules, multi-sensor corroboration, and automatic inhibition under ambiguous conditions. Operators often lack unlimited time; the legal question is whether the state and command structure designed the system and its operational environment so as to preserve reasonable opportunities for verification before lethal force was used.

Automation bias has direct legal significance in these circumstances. It describes the tendency of operators to over-rely on automated outputs or to treat them as more authoritative than competing information, a tendency intensified in military settings by stress, workload, hierarchy, uncertainty, and the perceived technical superiority of the system. Wood's category of will-offloading captures its legal meaning: the human appears to decide while substantive judgment has migrated to the system's output.[55] Saxon's analysis of speed reinforces it, since an operator with seconds to interpret a complex display exercises judgment only if interface, training, and doctrine permit something other than reflexive confirmation.[56]

These difficulties complicate any simple reliance on human-in-the-loop arrangements as proof of meaningful control. Such an architecture may improve on full automation in some contexts without guaranteeing meaningful control. A human may be asked to approve an engagement without understanding why the system classified a track as hostile; a supervisor may hold override power without the time to use it; a commander may authorize automatic mode without appreciating how the system behaves under rare or degraded conditions. Responsibility cannot be assessed by asking whether a human was present, but by asking whether the human role was epistemically and operationally capable of carrying legal responsibility.[57]

This perspective also helps distinguish individual error from institutional failure. Operators may bear responsibility where they ignore clear warnings, violate rules of engagement, or act recklessly. They should not become the default bearers of responsibility for a system whose interface, doctrine, and tempo render independent judgment practically impossible. De Vries's distinction between apparent and real responsibility gaps applies[58]: what presents as operator failure may be an institutional failure to design a decision environment in which operators could understand, challenge, or interrupt automated outputs. The inquiry must then move upward and outward — to training, doctrine, procurement, interface design, rules of engagement, legal review, and command authorization.

Air defense is central to the article because it reveals a gap between the responsibility assigned to operators and their practical capacity to exercise judgment. Responsibility gaps arise not only from the absence of a human decision-maker but from a mismatch between responsibility and capacity. A legal system may assign responsibility to an operator who lacks the time, information, or authority to make a legally meaningful decision. A commander may authorize a defensive mode without understanding the system's classification errors in complex environments. A state may claim that humans remain in control while its operational design makes that control largely symbolic. The gap opens between legal responsibility and the operational conditions required to exercise it.

Barbosa’s analysis locates state responsibility in the earlier decisions to procure, test, review, and authorize automated defensive systems, as well as in the circumstances of the eventual engagement.[59] Where a state knows that a system misclassifies in certain environments, or that

operators tend to defer to its outputs, continued deployment without mitigation becomes legally significant. Due diligence here demands more than verifying performance under ideal conditions; it demands testing under degraded, adversarial, ambiguous, and high-stress conditions, since those are the conditions in which air-defense systems are actually used. Article 36 review takes on a correspondingly specific meaning: it must examine how the system classifies threats, what uncertainty indicators it presents, whether operators can understand the basis of a classification, whether the system can be placed in different modes, whether automatic mode is restricted to specific environments, and whether logs permit later reconstruction. A review treating the system as merely a defensive technical platform misses the legal significance of human-machine interaction, since the method of warfare includes the decision architecture through which targets are identified and engaged.

Particular attention must also be paid to interface design.[60] *A* control architecture that ignores it may preserve formal legal standards while producing operational behavior that undermines them. Human control depends on how the system presents the world to the human, and not only on where the human sits in the loop. It depends equally on environment: a system acceptable in a closed military zone against incoming rockets may be unacceptable, or require different safeguards, in mixed civilian airspace, which vindicates Taddeo and Blanchard's insistence that purpose of use and human control are essential to any workable definition.[61]

Taken together, the three cases reveal distinct forms of weakened responsibility: Gaza through compressed review, Libya/Kargu-2 through autonomous engagement, and air defense through speed and automation bias. All three converge on the same insight: the legal problem is the degradation of the conditions that make human judgment and responsibility meaningful, and not autonomy as such. International law should therefore ask whether the human role was capable of performing the legal function assigned to it, rather than asking only whether a human was in, on, or out of the loop. Where speed, interface, and classification logic make independent judgment unrealistic, human presence is a weak basis for accountability, and part of the legal burden must be relocated to prior design, review, testing, authorization, and operational limitation.

Section V: Reconstructing Responsibility through the Control Envelope

The preceding cases show that algorithmic mediation weakens responsibility in different ways without displacing the applicable legal frameworks. This section reconstructs responsibility across individual, state, and private-supplier roles, then develops the control envelope as an operational standard linking foreseeability, traceability, reviewability, and meaningful human control throughout the system's life cycle.

A) Fractures rather than a void

The three cases reveal a set of fractures in the ordinary pathways through which law connects conduct, knowledge, control, and harm. Algorithmic classification precedes human approval and compresses review; a system continues the targeting process after activation and complicates attribution; formal human presence loses legal significance under speed, interface design, and automation bias. The common thread is a weakening of the conditions that allow law to identify who knew, who controlled, who could intervene, and who accepted the risk.

The vocabulary of the "responsibility gap" misleads if it implies a legal void.[62] *The* hard question is whether these frameworks remain practically usable once algorithmic systems distribute knowledge and control across multiple actors and technical processes.

B) Individuals: commanders, operators, technical personnel

Individual criminal responsibility remains essential, although it cannot by itself address every institutional failure associated with algorithmic systems. In a simple strike the relationship between commander, operator, target, weapon, and harm can be reconstructed from orders, communications, and battlefield evidence. In algorithmic targeting that relationship is mediated by model outputs, target profiles, system settings, training data, interface cues, and institutional tempo. The existing rules of criminal responsibility remain applicable, although their application requires closer attention to how authority, knowledge, and control are distributed. The question is whether a particular individual possessed the authority, knowledge, and capacity to prevent or shape the harmful outcome. A commander who authorizes a system despite known defects, an operator who knowingly ignores contrary information, and a technical officer who accepts a substantial risk of unlawful targeting all remain within established frameworks.[63]

Command responsibility remains important, but its application requires particular care in algorithmically mediated operations. Its requirements — effective control, knowledge or reason to know, failure to take necessary and reasonable measures — are complicated rather than removed by algorithmic systems. In Gaza-type targeting, a commander's responsibility may turn less on personal review of each file than on authorization of a workflow that predictably converts machine-generated classifications into rapid approvals. In Kargu-2-type scenarios it may turn on authorizing autonomous mode in an environment where lawful discrimination could not reasonably be expected. In air defense it may turn on permitting automatic or semi-automatic engagement in ambiguous airspace despite known misclassification risks. What matters is the commander's role in creating or tolerating the decision architecture that produced the harm, rather than proximity to the final engagement.[64]

Operators are especially vulnerable to blame because they are often the most visible human participants in the process. Yet visibility should not be mistaken for responsibility. An analyst in AI-assisted targeting receives outputs that have already framed the target as lawful or high-priority; an air-defense operator may have seconds to accept or reject a threat classification; the operator of a loitering munition may have launched a system whose later selection behavior he does not control. Responsibility may attach where the operator had meaningful information, time, authority, and a realistic opportunity to intervene. Where system design or command doctrine removed those conditions, blaming the operator converts a structural failure into an individual scapegoat.[65]

The mental element is therefore central to the assessment of individual criminal responsibility. International criminal law requires more than causal association with harm — intent, knowledge, or, depending on the offense and interpretive approach, awareness of a substantial risk — and algorithmic systems make the mental element harder to prove by separating those who design the system from those who authorize the mission and those who approve the target. Although proof may be difficult, the mental element can still be established where the available evidence shows knowledge or conscious acceptance of a substantial risk. Dolus eventualis and recklessness reach the person who knows of a substantial risk that the system will violate IHL and nevertheless accepts or disregards it[66], while Kwik's work denies protection to ignorance that is cultivated, engineered, or relied upon as part of negligent institutional practice.[67]

The legal position of programmers, technical personnel, and system validators is more difficult to determine.[68] *Their* role cannot be ignored either. A developer who knowingly encodes an unlawful target profile, suppresses known error rates, validates a model under misleading conditions, or fails to report defects likely to produce unlawful attacks may attract individual

responsibility depending on knowledge, contribution, and causal connection. The right approach asks whether the technical actor knowingly contributed to a system whose unlawful use was sufficiently foreseeable, rather than expanding criminal liability mechanically.[69]

C) The state and its private suppliers

These limits make state responsibility an indispensable part of the legal response. The state procures, reviews, authorizes, deploys, constrains, monitors, and withdraws military systems, and it may bear responsibility for an internationally wrongful act attributable to it even where no individual can be convicted. Barbosa shows that AWS challenges the human-action assumptions of traditional attribution without removing the state from the legal picture.[70] A state cannot enjoy algorithmic speed, scale, and opacity while treating harmful outcomes as products of an autonomous technical sphere beyond legal responsibility.

Attribution must therefore be reconstructed institutionally. In Gaza-type settings, state conduct includes adopting an AI-assisted targeting workflow, defining target profiles, validating outputs, setting review standards, and authorizing operational tempo. In Kargu-2-type settings it includes deploying a loitering munition with autonomous search and engagement functions under particular environmental constraints. In air-defense settings it includes activating automated modes, defining threat categories, training operators, and permitting engagement under uncertainty. The system's output is examined as the consequence of state-controlled operational architecture rather than as an independent legal act by a machine.[71]

Due diligence bridges state responsibility and operational control. A state deploying autonomous or AI-enabled systems must take reasonable steps to understand their limits, test them under realistic conditions, restrict their use to appropriate environments, train operators, monitor performance, and preserve records. The standard is feasible precaution and responsible control over means and methods of warfare, not absolute safety, since international law demands no impossible certainty in war. Barbosa's "knew or should have known" formulation links responsibility to the state's duty to build knowledge before deployment.[72] That duty also forecloses the misuse of unpredictability as a defense: a state that deploys a system precisely because it operates adaptively or autonomously under uncertain conditions cannot then invoke the unforeseeability of its behavior, and force majeure should not become a convenient escape where the state chose a system with known opacity.[73]

Private suppliers add a further layer. AI-enabled systems depend on proprietary models, sensor packages, cloud services, training datasets, software updates, and contractor support — technically decisive influences that traditional battlefield responsibility captures poorly. International law offers no general regime of corporate criminal responsibility for such cases, and private companies are no substitute for state responsibility. Proprietary restrictions cannot relieve a state of its duty to understand, test, audit, and constrain a system before use. A state unable to do so has not established the conditions required for lawful deployment.[74] *For* present purposes the corporate question matters chiefly because black-box systems, undisclosed datasets, and inaccessible logs weaken both IHL compliance and post-incident accountability.

The institutional distribution of responsibility varies across the three cases. Gaza directs attention to designers of target profiles, intelligence authorities, commanders who set approval tempo, and officials who decide whether machine-generated outputs may structure lethal decisions. Libya/Kargu-2 directs it to those who select autonomous mode, define the mission envelope, approve deployment conditions, and preserve engagement logs. Air defense directs it to those who

design interfaces, train operators, authorize automatic modes, and judge when civilian or friendly presence makes automation too risky. Responsibility runs along a chain in each configuration, and the task of law is to identify which links carried knowledge, authority, and capacity.

This distribution also separates criminal responsibility from accountability more broadly. Criminal responsibility should remain tied to individual culpability and should not dissolve into mere association with a system. State responsibility can address institutional failures where criminal culpability cannot be proved, and administrative, disciplinary, procurement, and review mechanisms can address failures below the criminal threshold that nonetheless undermine compliance. Crootof's "war torts" proposal registers exactly this dissatisfaction with a framework that either convicts an individual or leaves victims without remedy.[75] Addressing algorithmic harm therefore requires several complementary forms of accountability, including individual, state, administrative, and institutional responsibility.[76]

D) The control envelope as an operational standard

The legal inquiry should begin before deployment and continue after any harmful incident. Who defined the system's function? Who approved the target profile? Who validated the model? Who set the review time? Who authorized the mode of operation? Who preserved the logs, monitored performance, and held authority to suspend use? These questions do not replace existing doctrines; they make them usable, allowing criminal responsibility to focus on culpable knowledge and control, state responsibility on institutional failure, and corporate involvement on traceability, due diligence, and contribution.

The proposed control envelope provides a framework for answering these questions: the documented operational, technical, and legal boundaries within which an autonomous or AI-enabled military system may be deployed with a reasonable degree of foreseeability, traceability, and human accountability.[77] *It* encompasses the legal function assigned to the system, the target category, the environment, the duration of operation, the permissible degree of autonomy, the human role, the review process, the data inputs, the classification thresholds, the known error profile, the audit trail, and the conditions for suspension or withdrawal. It serves both a preventive and an accountability function: before deployment, it identifies the conditions of lawful use; after an incident, it provides the information needed to reconstruct responsibility.

The concept builds on and modifies existing accounts of meaningful human control.[78] *It* integrates these insights and adds a requirement of its own: control may be distributed across design, review, authorization, operation, and monitoring, but the distribution must be documented, legally intelligible, and capable of sustaining responsibility. Its operational content can be organized around ten connected elements.

Functional specification. The state must define what the system may do — collect intelligence, classify objects, rank candidates, estimate collateral harm, select, track, or engage — since the legal consequences differ sharply across these functions. Definitions vary precisely in the weight they place on autonomy, adaptive capability, human control, and purpose of use, which is why vague labels must be refused.[79]

Target definition. No system should operate on an undefined or legally ambiguous target category. The envelope must specify whether the system is directed at munitions, radar emitters, vehicles, vessels, buildings, persons, communication patterns, or combinations of indicators, and whether that category is legally stable or context-dependent. Since target profiles encode assumptions and

military priorities into technical criteria[80], those assumptions must remain visible enough to be reviewed, challenged, and revised.

Spatial and temporal limitation. Area of operation, duration of activation, permissible routes, engagement zones, no-strike areas, and deactivation conditions must be defined, most urgently for loitering and automated defensive systems. Narrow frames make it plausible that human judgment genuinely structured the system's operation in advance; broad and fluid ones place a heavier justificatory burden on the deploying authority.

Environmental suitability. A system adequate in controlled testing may fail amid civilians, camouflage, damaged infrastructure, spoofing, degraded sensors, mixed traffic, surrendering persons, or communication loss.[81] The envelope must specify where a system may *not* be used as clearly as where it may, since legality depends on the match between system capacity and operational context.

Data and classification governance. For AI-enabled systems, data sources, model assumptions, confidence thresholds, error rates, and validation procedures shape legal outcomes rather than forming technical background. The envelope must identify what data were used, how current they are, what confidence threshold applies, what error rates are known, and how uncertainty reaches reviewers. Scores, categories, and recommendations should remain sources of information for the legal decision-maker rather than substitutes for legal judgment.

Human review capacity. Human involvement must be specified functionally rather than symbolically: who reviews the output, what information is available, how much time, what authority to reject or delay, and what happens when uncertainty persists. A reviewer who cannot understand the basis of a recommendation, access contrary information, or resist institutional pressure cannot bear the legal weight assigned to "human control".[82]

Abort, override, and suspension authority. The envelope must define who can stop the system and on what conditions — real-time abort where feasible, and otherwise pre-planned self-deactivation, loss-of-link behavior, geofencing, no-engagement conditions, and mandatory suspension when outputs become anomalous. Where speed limits real-time intervention, more weight falls on prior constraints, mode selection, and automatic inhibition under uncertainty.

Traceability and logging. Responsibility cannot be reconstructed from a system that leaves no usable record. Logs must be sufficient to identify system mode, data inputs, sensor detections, classification outputs, confidence levels, human interactions, overrides, communication status, and reasons for engagement or non-engagement where technically feasible. A lack of traceability weakens the institutional basis of responsibility and creates more than a simple evidentiary difficulty.

Testing and review under realistic conditions. Article 36 review must go beyond a paper assessment of intended use to examine modes of operation, target profiles, human-machine interface, update mechanisms, cybersecurity vulnerabilities, error rates, environmental assumptions, and training requirements.[83] A state that reviews a system only under ideal conditions cannot later treat failure under foreseeable battlefield conditions as an unforeseeable accident.

Post-deployment monitoring and learning. The envelope is not fixed at the moment of legal review. It must be updated when the system behaves unexpectedly, when the environment changes, when intelligence assumptions fail, when operators show signs of overreliance, or when near-misses expose hidden risk. An unexpected initial failure may be difficult to prevent, but continued use after

warning signs have emerged raises a more serious question of institutional responsibility, since ignorance becomes legally significant when institutions preserve it despite opportunities to learn.[84]

E) What the envelope changes

These elements connect individual and state responsibility. For individuals, the envelope identifies who held knowledge, authority, and capacity — whether a commander authorized use outside approved conditions, an operator ignored clear warnings, a technical officer concealed known errors, or a reviewer approved a target despite unresolved uncertainty. For the state, it supplies a record of due diligence or its absence. For private suppliers, it fixes the minimum access a state must retain to review, test, audit, and constrain a system, so that proprietary opacity cannot substitute for legal traceability.[85]

The concept occupies a middle position between a categorical prohibition and an unqualified reliance on existing law. It requires no total ban on every system with autonomous or AI-enabled functions, since some may operate lawfully within narrow, well-tested, and well-documented conditions. It refuses equally the claim that existing law suffices merely because IHL formally applies, because law must be able to see the decision — to reconstruct the chain between human authorization, system output, and harmful effect.

Its application must be adapted to the function of the system and the environment in which it is used. In Gaza-type targeting it centres on target-profile design, data reliability, confidence thresholds, review time, uncertainty presentation, and resistance to rubber-stamping. In Kargu-2-type engagement it centres on mission boundaries, target class, mode selection, loss-of-link behavior, environmental suitability, abort rules, and engagement logs. In air defense it centres on threat categories, engagement zones, automatic-mode restrictions, operator interface, automation bias, civilian-airspace coordination, and uncertainty thresholds. The envelope is a legal-operational method for asking what control must mean in each configuration, which matters because autonomy is not binary: a fixed rule requiring human approval of every engagement would be too rigid for some defensive contexts and too weak for AI-assisted targeting where approval is merely formal.

The framework also changes the way foreseeability is assessed. In ordinary legal analysis it is assessed retrospectively — could the actor have anticipated the outcome? — whereas in algorithmic warfare it must be built beforehand, through testing, documentation, constraints, monitoring, and the preservation of uncertainty. A state that failed to examine a system under realistic conditions, ignored known error rates, used vague target profiles, or allowed operators to approve outputs too quickly to understand them cannot claim that the behavior was unforeseeable. Foreseeability depends partly on the knowledge and precautions developed by the institution before deployment.

Traceability likewise becomes a condition of lawful and accountable use rather than merely a tool for later investigation. Where a system cannot produce or preserve records sufficient to reconstruct its outputs, modes, and human interactions, the state has chosen a method of warfare that weakens accountability by design - a choice that should bear on state responsibility, weapons review, evidentiary presumptions, and the assessment of due diligence. The legal system should not reward architectures that make responsibility impossible to prove.

The framework also treats meaningful human control as a process distributed across several stages and actors. Control is neither a single moment nor a single actor, but a chain of human judgment distributed across design, legal review, command authorization, operator interaction, monitoring, and post-incident assessment. Viewing control as distributed helps identify the

responsibility attached to each stage of design, authorization, operation, and review, by asking which actor controlled which part of the chain, and it forestalls the common error of assigning everything either to the final operator or to an abstract "system." The machine is not the responsible agent; the legally relevant question is how human institutions structured its operation.

The control envelope should therefore serve as a minimum standard for determining whether a system may be used lawfully and accountably in a particular environment. Where no envelope can be built for a particular system in a particular environment, the system should not be used there — a position narrower than a categorical ban and stronger than a general call for human oversight. Deployment is lawful only where the state can show that operation remains within documented boundaries capable of sustaining IHL compliance and responsibility. Where behavior cannot reasonably be predicted, outputs cannot be traced, human review is symbolic, or no actor can explain the basis for engagement, the conditions of lawful accountability are absent.

Conclusion:

Autonomous and AI-enabled military systems do not create a single or uniform responsibility gap. The three cases examined in this article reveal different ways in which algorithmic mediation affects responsibility. In Gaza, AI-assisted targeting may compress human review through algorithmic classification and target profiling. In Libya/Kargu-2, autonomous engagement may complicate foreseeability and attribution after activation. In air-defense systems, formal human presence may remain in place while speed, interface design, and automation bias weaken meaningful judgment. Across all three cases, the relevant rules of international law remain applicable. International humanitarian law, individual criminal responsibility, command responsibility, state responsibility, and weapons review obligations continue to govern these systems.

The main difficulty lies in applying these rules when the information and institutional safeguards needed to establish responsibility are incomplete or poorly documented. Responsibility depends on the ability to reconstruct who authorized the system, what risks were known, what limits were imposed, how the system produced its output, and whether human judgment remained capable of carrying legal responsibility. For this reason, the article proposes the control envelope as a documented set of operational, technical, and legal boundaries within which an autonomous or AI-enabled military system may be used with a reasonable degree of foreseeability, traceability, reviewability, and human accountability.

First: Key Findings

Four principal findings emerge from the analysis. First, the responsibility gap in algorithmic warfare is not a complete legal void. It is better understood as a weakening of the practical conditions that allow the law to connect conduct, knowledge, control, and harm. Second, the presence of a human in or on the loop is insufficient to establish meaningful control. What matters is whether the human actor had the time, information, authority, training, and institutional freedom needed to exercise legal judgment. Third, responsibility cannot be located only at the final moment of approval or engagement. It must be traced across the wider institutional chain, including design, data selection, target-profile construction, system validation, legal review, command authorization, operator interaction, logging, and post-deployment monitoring. Fourth, state responsibility becomes central because the state controls the conditions under which these systems are procured, reviewed, tested, used, constrained, monitored, and withdrawn.

Taken together, these findings clarify the role of the proposed control envelope. The legality of using autonomous or AI-enabled systems should not be assessed solely by asking whether the system is autonomous or whether a human was formally present. The inquiry should determine whether the system operated within documented limits that made its behavior reasonably foreseeable, its outputs traceable, its human review meaningful, and its effects legally attributable. Where such an envelope exists and is respected, existing legal frameworks can operate more effectively. Where the required boundaries cannot be defined and documented, lawful and accountable use cannot be assured.

Second: Recommendations

States should treat the control envelope as a minimum condition for the lawful and accountable use of autonomous and AI-enabled military systems. Before operational use, the responsible authority should define the system's function, target category, operational area, time limits, data sources, classification thresholds, known error rates, human review procedures, abort mechanisms, logging duties, and post-deployment monitoring requirements.

Article 36 weapons review should also be adapted to the technical and operational features of algorithmic warfare. Review should not be limited to the physical platform or the weapon that produces the kinetic effect. It should also examine AI-assisted targeting tools, decision-support systems, target profiles, training data, human-machine interfaces, system modes, cybersecurity vulnerabilities, environmental assumptions, operator workload, auditability, and the possibility of suspending the system when its outputs become unreliable.

Human involvement should be assessed according to its practical quality rather than the formal presence of an operator or reviewer. Human review must be supported by sufficient time, accessible information, visible uncertainty, authority to reject or delay, and an institutional culture that permits refusal. A human signature or approval click should not be treated as legally meaningful where the surrounding process makes independent judgment unrealistic.

Traceability should form part of the legal and technical conditions established before operational use. Systems should generate and retain records that allow the later reconstruction of inputs, outputs, confidence scores, operating modes, human interactions, override opportunities, communication status, and the reasons for engagement or non-engagement where technically feasible. A system that cannot be traced may make accountability practically impossible and should not be used in contexts where decisions involving lethal force require legal judgment.

Finally, systems that cannot be contained within a robust control envelope should be prohibited or strictly restricted in the relevant operational environment, especially where they would be used in civilian-dense areas, rely on broad or ambiguous target profiles, allow only compressed review time, produce opaque classifications, or fail to preserve adequate records. The decisive issue is whether human institutions have established and preserved the conditions required for legal judgment and responsibility throughout the system's operational life cycle.

Citations:

[1] Dan Saxon, *Fighting Machines: Autonomous Weapons and Human Dignity* (Philadelphia: University of Pennsylvania Press, 2022), pp. 42–46; Nathan Wood, "Autonomous and AI-Enabled Systems: Extensions or Replacements of Human Will and Control?," *Ethics and Information Technology* 28, no. 3 (2026): 9–12.

[2] Rebecca Crootof, "The Killer Robots Are Here: Legal and Policy Implications," *Cardozo Law Review* 36, no. 5 (2015): 1837–1915; Rebecca Crootof, "War Torts: Accountability for Autonomous Weapons," *University of Pennsylvania Law Review* 164, no. 6 (2016): 1347–1402.

[3] Barry de Vries, *Individual Criminal Responsibility for Autonomous Weapons Systems in International Criminal Law*, International Humanitarian Law Series 65 (Leiden: Brill Nijhoff, 2023), p. 123; Jonathan Kwik, "The Conceptual Roots of the Criminal Responsibility Gap in Autonomous Weapon Systems," *Melbourne Journal of International Law* 24 (2023): 4.

[4] Daniele Amoroso and Guglielmo Tamburrini, "Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues," *Current Robotics Reports* 1, no. 4 (2020): 187–194; Lutiana Valadares Fernandes Barbosa, *Autonomous Weapons Systems and the Responsibility of States: Challenges and Possibilities* (Boca Raton: CRC Press, 2025), pp. 102–119; Berenike Prem, "Control-by-Design? Autonomous Weapons Systems as Technopolitical Projects," *Contemporary Security Policy* 47, no. 2 (2026): 415–443.

[5] Afonso Seixas-Nunes, *The Legality and Accountability of Autonomous Weapon Systems: A Humanitarian Law Perspective* (Cambridge: Cambridge University Press, 2022), pp. 127–128; Mariarosaria Taddeo and Alexander Blanchard, "A Comparative Analysis of the Definitions of Autonomous Weapons Systems," *Science and Engineering Ethics* 28, art. 37 (2022): 10–18.

[6] The envelope therefore stands close to, but not within, the meaningful-human-control literature: where that literature asks what quality of human involvement the law should demand, the envelope asks the prior institutional question of what a state must document, constrain, and preserve before any human involvement can bear legal weight at all.

[7] Sparrow's concern went beyond the observation that machines lack moral agency: the ordinary structure of responsibility may itself be disrupted when lethal outcomes emerge from systems acting beyond direct human decision. For more on this topic, see Robert Sparrow, "Killer Robots," *Journal of Applied Philosophy* 24, no. 1 (2007): 62–77.

[8] Crootof later proposed that criminal law alone may be insufficient, so that a supplementary regime — "war torts" — might be required (Rebecca Crootof, "War Torts," pp. 1347–1402). For more on this topic, see Rebecca Crootof, "The Killer Robots Are Here," pp. 1837–1915.

[9] Barry de Vries, *Individual Criminal Responsibility*, p. 123.

[10] Jonathan Kwik, "The Conceptual Roots of the Criminal Responsibility Gap," p. 4.

[11] Jonathan Kwik, "'I Plead Ignorance': Autonomous Weapons and Criminal Liability for Not Knowing the Knowable," *International Law Studies* 107 (2026): 256–290.

[12] International Committee of the Red Cross, *Autonomous Weapon Systems and International Humanitarian Law: Selected Issues* (Geneva: International Committee of the Red Cross, 2025).

[13] Daniele Amoroso and Guglielmo Tamburrini, "Autonomous Weapons Systems and Meaningful Human Control," pp. 187–194.

[14] Dan Saxon, *Fighting Machines*, pp. 3–4, 42–46.

[15] Nathan Wood, "Autonomous and AI-Enabled Systems," pp. 9–12.

[16] Afonso Seixas-Nunes, *The Legality and Accountability of Autonomous Weapon Systems*, pp. 127–128.

[17] Ibid., pp. 214–225.

[18] Ibid., pp. 56–57.

[19] Lutiana Valadares Fernandes Barbosa, *Autonomous Weapons Systems and the Responsibility of States*, p. ix.

[20] Ibid., pp. 140–145.

[21] Ibid., pp. 102–119, 140–166.

[22] A narrow definition excludes systems that never fire yet decisively structure targeting decisions; an overly broad one collapses the differences between AWS, LAWS, automated defensive systems, and AI-enabled

decision-support tools. For more on this topic, see Mariarosaria Taddeo and Alexander Blanchard, "A Comparative Analysis of the Definitions of Autonomous Weapons Systems," pp. 10–18.

[23] Berenike Prem, "Control-by-Design?," pp. 415–443.

[24] Ingvild Bode and Hendrik Huelss, *Autonomous Weapons Systems and International Norms* (Montreal and Kingston: McGill-Queen's University Press, 2022); Nathan Wood, "Autonomous and AI-Enabled Systems," pp. 11–12.

[25] Neil Davison, "A Legal Perspective: Autonomous Weapon Systems under International Humanitarian Law," in *Perspectives on Lethal Autonomous Weapon Systems*, UNODA Occasional Papers No. 30 (New York: United Nations Office for Disarmament Affairs, 2017), pp. 5, 7, 9; International Committee of the Red Cross, *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons* (Geneva: International Committee of the Red Cross, 2016).

[26] Afonso Seixas-Nunes, *The Legality and Accountability of Autonomous Weapon Systems*, pp. 5–6.

[27] Nathan Wood, "Autonomous and AI-Enabled Systems," pp. 9–12.

[28] A human in the loop may lack time, information, training, or authority to question the system. A human on the loop may hold a theoretical power to abort while being unable to detect error quickly enough to use it. A system operating without contemporaneous intervention may nonetheless be constrained by a narrow mission, target type, area, and temporal window.

[29] Dan Saxon, *Fighting Machines*, pp. 42–46.

[30] Berenike Prem, "Control-by-Design?," pp. 415–443.

[31] Lutiana Valadares Fernandes Barbosa, *Autonomous Weapons Systems and the Responsibility of States*, pp. 102–119.

[32] The CCW/GGE debate registers the same unresolved tension. States have repeatedly affirmed that IHL applies fully to AWS and that humans must retain responsibility or judgment in the use of force, while disagreeing over whether to frame the requirement as meaningful human control, appropriate human judgment, human-machine interaction, or context-specific control. The vocabulary matters — "control" suggests an operational relationship between human and system, "judgment" emphasizes the quality of human legal and ethical assessment — though the underlying challenge is unchanged: ensuring that human involvement is not merely symbolic (Nathan Wood, "Autonomous and AI-Enabled Systems," pp. 3–4).

[33] The Israeli military has disputed the implication that such systems autonomously determine targets, stating that they assist analysts and do not constitute the sole basis for approval (Vera Bergengruen, "How Israel Uses AI in Gaza—and What It Might Mean for the Future of Warfare," *Time*, December 18, 2024, https://time.com/7202584/gaza-ukraine-ai-warfare/). For more on this topic, see Harry Davies, Bethan McKernan, and Dan Sabbagh, "'The Machine Did It Coldly': Israel Used AI to Identify 37,000 Hamas Targets," *The Guardian*, April 3, 2024, https://www.theguardian.com/world/2024/apr/03/israel-gaza-ai-database-hamas-airstrikes.

[34] Nathan Wood, "Autonomous and AI-Enabled Systems," pp. 9–12.

[35] Berenike Prem, "Control-by-Design?," pp. 415–443.

[36] Dan Saxon, *Fighting Machines*, pp. 42–46.

[37] Nathan Wood, "Autonomous and AI-Enabled Systems," pp. 11–12.

[38] Afonso Seixas-Nunes, *The Legality and Accountability of Autonomous Weapon Systems*, pp. 5–6.

[39] Lutiana Valadares Fernandes Barbosa, *Autonomous Weapons Systems and the Responsibility of States*, pp. 102–119.

[40] The phrasing became central to public and scholarly debate and should be handled with care, since the report offers no full forensic reconstruction of a specific strike and settles neither the mode of operation, nor human supervision, nor casualty causation. See also United Nations Security Council, *Final Report of the Panel of Experts on Libya Established Pursuant to Security Council Resolution 1973 (2011)*, S/2021/229, March 8, 2021, para. 63.

[41] Afonso Seixas-Nunes, *The Legality and Accountability of Autonomous Weapon Systems*, pp. 127–128.

[42] Nathan Wood, "Autonomous and AI-Enabled Systems," pp. 9–12.

[43] Jonathan Kwik, "The Conceptual Roots of the Criminal Responsibility Gap," p. 4.

[44] Lutiana Valadares Fernandes Barbosa, *Autonomous Weapons Systems and the Responsibility of States*, pp. 140–145.

[45] Ibid., pp. 102–119.

[46] Mariarosaria Taddeo and Alexander Blanchard, "A Comparative Analysis of the Definitions of Autonomous Weapons Systems," pp. 10–18.

[47] Afonso Seixas-Nunes, *The Legality and Accountability of Autonomous Weapon Systems*, pp. 214–225.

[48] Lutiana Valadares Fernandes Barbosa, *Autonomous Weapons Systems and the Responsibility of States*, pp. 140–166.

[49] Berenike Prem, "Control-by-Design?," pp. 415–443.

[50] Dan Saxon, *Fighting Machines*, pp. 8–13; Ingvild Bode and Hendrik Huelss, *Autonomous Weapons Systems and International Norms*.

[51] Dan Saxon, *Fighting Machines*, pp. 8–10.

[52] Nathan Wood, "Autonomous and AI-Enabled Systems," p. 10.

[53] Ibid., pp. 9–12.

[54] The downing of Iran Air Flight 655 by USS Vincennes in 1988 is commonly analysed as a case of misidentification, stress, and failed situational awareness in a technologically mediated combat environment (United States Department of Defense, *Formal Investigation into the Circumstances Surrounding the Downing of Iran Air Flight 655 on 3 July 1988*, prepared by Rear Admiral William M. Fogarty [Washington, DC: U.S. Department of Defense, 1988], https://apps.dtic.mil/sti/pdfs/ADA203577.pdf). Patriot fratricide incidents during the 2003 Iraq War have been discussed in relation to automation, classification, operator trust, and the difficulty of challenging system outputs under pressure (Defense Science Board Task Force, *Patriot System Performance: Report Summary* [Washington, DC: Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics, 2005], https://stib.cto.mil/wp-content/uploads/reports/2000s/ADA435837.pdf). Neither is an AWS case in the strict sense.

[55] Nathan Wood, "Autonomous and AI-Enabled Systems," pp. 11–12.

[56] Dan Saxon, *Fighting Machines*, pp. 42–46.

[57] Daniele Amoroso and Guglielmo Tamburrini, "Autonomous Weapons Systems and Meaningful Human Control," pp. 187–194.

[58] Barry de Vries, *Individual Criminal Responsibility*, p. 123.

[59] Lutiana Valadares Fernandes Barbosa, *Autonomous Weapons Systems and the Responsibility of States*, pp. 102–119.

[60] Many automated systems present outputs in binary form — hostile or friendly, engage or do not engage, threat or non-threat — whereas legal judgment turns on uncertainty, context, and competing indicators. Where an interface hides uncertainty, suppresses contrary indicators, or generates pressure for immediate confirmation, the law's demand for verification becomes difficult to satisfy in practice.

[61] Mariarosaria Taddeo and Alexander Blanchard, "A Comparative Analysis of the Definitions of Autonomous Weapons Systems," pp. 10–18.

[62] Individuals may incur criminal responsibility for war crimes where the material and mental elements are satisfied. Commanders may be responsible for crimes committed by forces under their effective command and control where they knew or should have known and failed to prevent or punish. States may be responsible for internationally wrongful acts attributable to them, and must review new weapons, means, and methods of warfare (Rome Statute of the International Criminal Court, July 17, 1998, 2187 U.N.T.S. 90, arts. 25, 28, 30; International Law Commission, "Draft Articles on Responsibility of States for Internationally Wrongful Acts, with Commentaries," in *Yearbook of the International Law Commission, 2001*, vol. 2, part 2

[New York: United Nations, 2001], arts. 4, 8, 16; Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts [Additional Protocol I], June 8, 1977, 1125 U.N.T.S. 3, art. 36).

[63] Afonso Seixas-Nunes, *The Legality and Accountability of Autonomous Weapon Systems*, pp. 214–225.

[64] Rome Statute, art. 28; Afonso Seixas-Nunes, *The Legality and Accountability of Autonomous Weapon Systems*, pp. 56–57.

[65] Barry de Vries, *Individual Criminal Responsibility*, p. 123; Dan Saxon, *Fighting Machines*, pp. 83–84.

[66] Afonso Seixas-Nunes, *The Legality and Accountability of Autonomous Weapon Systems*, pp. 214–225.

[67] Jonathan Kwik, "I Plead Ignorance."

[68] They are not combatants making battlefield targeting decisions, and their work occurs long before deployment and far from the operational context, so treating technical contribution as automatically equivalent to criminal participation would be legally dangerous.

[69] Legal advisers and weapons reviewers belong to the architecture as well: a review approving a system on the basis of incomplete testing, vague target profiles, or unrealistic assumptions may create no criminal responsibility, yet it bears on state responsibility and on what commanders and officials knew, and a review that ignores autonomous modes, data dependencies, interface, error rates, environment, and traceability helps produce the very opacity later described as a responsibility gap (International Committee 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* [Geneva: International Committee of the Red Cross, 2006]; Lutiana Valadares Fernandes Barbosa, *Autonomous Weapons Systems and the Responsibility of States*, pp. 102–119). For more on this topic, see Afonso Seixas-Nunes, *The Legality and Accountability of Autonomous Weapon Systems*, pp. 214–225; Berenike Prem, "Control-by-Design?," pp. 415–443.

[70] Lutiana Valadares Fernandes Barbosa, *Autonomous Weapons Systems and the Responsibility of States*, pp. 140–145.

[71] Ibid., pp. 140–166.

[72] Ibid., pp. 111–119.

[73] Ibid., pp. 80–83, 157.

[74] Corporate accountability may still arise through domestic criminal or civil law, export controls, procurement rules, contractual obligations, and human-rights due diligence, and international criminal responsibility may reach individuals within companies who knowingly aid or assist international crimes. For more on this topic, see ibid., pp. 143–145, 166.

[75] Rebecca Crootof, "War Torts," pp. 1347–1402.

[76] Two extremes must accordingly be rejected: technological fatalism, which understates the continuing force of IHL, Article 36 review, individual criminal responsibility, state responsibility, and due diligence; and legal formalism, which overlooks the practical conditions without which legal rules cannot be enforced.

[77] The term should be distinguished from a technical safety envelope. In engineering, an envelope describes the conditions under which a system operates safely or reliably; the legal envelope proposed here includes technical reliability while extending well beyond it.

[78] Amoroso and Tamburrini already show that control cannot be reduced to pressing a button (Daniele Amoroso and Guglielmo Tamburrini, "Autonomous Weapons Systems and Meaningful Human Control," pp. 187–194); Saxon shows that system speed can empty human presence of normative content (Dan Saxon, *Fighting Machines*, pp. 42–46); Wood shows that contemporaneous supervision is not always required, provided ex ante constraints allow the system to extend rather than offload human will (Nathan Wood, "Autonomous and AI-Enabled Systems," pp. 9–12).

[79] Mariarosaria Taddeo and Alexander Blanchard, "A Comparative Analysis of the Definitions of Autonomous Weapons Systems," pp. 10–18.

[80] Berenike Prem, "Control-by-Design?," pp. 415–443.

[81] Ibid.

[82] Dan Saxon, *Fighting Machines*, pp. 42–46; Nathan Wood, "Autonomous and AI-Enabled Systems," pp. 11–12.

[83] Lutiana Valadares Fernandes Barbosa, *Autonomous Weapons Systems and the Responsibility of States*, pp. 102–119.

[84] Jonathan Kwik, "I Plead Ignorance."

[85] Lutiana Valadares Fernandes Barbosa, *Autonomous Weapons Systems and the Responsibility of States*, pp. 111–119, 140–166.

Bibliography:

A. Legal Instruments and International Legal Materials

- International Law Commission. "Draft Articles on Responsibility of States for Internationally Wrongful Acts, with Commentaries." In *Yearbook of the International Law Commission, 2001*, vol. 2, part 2. New York: United Nations, 2001. https://legal.un.org/ilc/texts/instruments/english/commentaries/9_6_2001.pdf.
- Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts. June 8, 1977. 1125 U.N.T.S. 3. https://www.ohchr.org/en/instruments-mechanisms/instruments/protocol-additional-geneva-conventions-12-august-1949-and.
- Rome Statute of the International Criminal Court. July 17, 1998. 2187 U.N.T.S. 90. https://www.icc-cpi.int/sites/default/files/NR/rdonlyres/ADD16852-AEE9-4757-ABE7-9CDC7CF02886/283503/RomeStatutEng1.pdf.

B. Official Reports and Policy Documents

- Davison, Neil. "A Legal Perspective: Autonomous Weapon Systems under International Humanitarian Law." In *Perspectives on Lethal Autonomous Weapon Systems*, UNODA Occasional Papers No. 30, 5–18. New York: United Nations Office for Disarmament Affairs, 2017. https://www.icrc.org/sites/default/files/document/file_list/autonomous_weapon_systems_under_international_humanitarian_law.pdf.
- Defense Science Board Task Force. *Patriot System Performance: Report Summary*. Washington, DC: Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics, January 2005. https://stib.cto.mil/wp-content/uploads/reports/2000s/ADA435837.pdf.
- Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems. *Report of the 2019 Session*. CCW/GGE.1/2019/3. Geneva: United Nations, 2019. https://documents.unoda.org/wp-content/uploads/2020/09/CCW_GGE.1_2019_3_E.pdf.
- International Committee 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*. Geneva: International Committee of the Red Cross, 2006. https://library.icrc.org/library/docs/DOC/icrc_002_0902.pdf.
- International Committee of the Red Cross. *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons*. Geneva: International Committee of the Red Cross, 2016. https://www.icrc.org/en/publication/4283-autonomous-weapons-systems.

- International Committee of the Red Cross. *Autonomous Weapon Systems and International Humanitarian Law: Selected Issues*. Geneva: International Committee of the Red Cross, 2025. https://www.icrc.org/en/article/autonomous-weapon-systems-and-international-humanitarian-law-selected-issues.
- United Nations Security Council. *Final Report of the Panel of Experts on Libya Established Pursuant to Security Council Resolution 1973 (2011)*. S/2021/229. March 8, 2021. https://www.undocs.org/S/2012/163
- United States Department of Defense. *Formal Investigation into the Circumstances Surrounding the Downing of Iran Air Flight 655 on 3 July 1988*. Prepared by Rear Admiral William M. Fogarty. Washington, DC: U.S. Department of Defense, 1988. https://apps.dtic.mil/sti/pdfs/ADA203577.pdf.

C. Books

- Barbosa, Lutiana Valadares Fernandes. *Autonomous Weapons Systems and the Responsibility of States: Challenges and Possibilities*. Boca Raton: CRC Press, 2025.
- Bode, Ingvild, and Hendrik Huelss. *Autonomous Weapons Systems and International Norms*. Montreal and Kingston: McGill-Queen's University Press, 2022.
- de Vries, Barry. *Individual Criminal Responsibility for Autonomous Weapons Systems in International Criminal Law*. International Humanitarian Law Series 65. Leiden: Brill Nijhoff, 2023.
- Saxon, Dan. *Fighting Machines: Autonomous Weapons and Human Dignity*. Philadelphia: University of Pennsylvania Press, 2022.
- Seixas-Nunes, Afonso. *The Legality and Accountability of Autonomous Weapon Systems: A Humanitarian Law Perspective*. Cambridge: Cambridge University Press, 2022.

D. Journal Articles :

- Amoroso, Daniele, and Guglielmo Tamburrini. "Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues." *Current Robotics Reports* 1, no. 4 (2020): 187–194.
- Crootof, Rebecca. "The Killer Robots Are Here: Legal and Policy Implications." *Cardozo Law Review* 36, no. 5 (2015): 1837–1915.
- Crootof, Rebecca. "War Torts: Accountability for Autonomous Weapons." *University of Pennsylvania Law Review* 164, no. 6 (2016): 1347–1402.
- Kwik, Jonathan. "The Conceptual Roots of the Criminal Responsibility Gap in Autonomous Weapon Systems." *Melbourne Journal of International Law* 24 (2023): 1–26.
- Kwik, Jonathan. "'I Plead Ignorance': Autonomous Weapons and Criminal Liability for Not Knowing the Knowable." *International Law Studies* 107 (2026): 256–290.
- Prem, Berenike. "Control-by-Design? Autonomous Weapons Systems as Technopolitical Projects." *Contemporary Security Policy* 47, no. 2 (2026): 415–443.
- Sparrow, Robert. "Killer Robots." *Journal of Applied Philosophy* 24, no. 1 (2007): 62–77.
- Taddeo, Mariarosaria, and Alexander Blanchard. "A Comparative Analysis of the Definitions of Autonomous Weapons Systems." *Science and Engineering Ethics* 28, article 37 (2022): 1–22.
- Wood, Nathan. "Autonomous and AI-Enabled Systems: Extensions or Replacements of Human Will and Control?" *Ethics and Information Technology* 28, article 3 (2026): 1–16.

E. Newspaper Articles:

- Abraham, Yuval. "'Lavender': The AI Machine Directing Israel's Bombing Spree in Gaza." *+972 Magazine* and *Local Call*, April 3, 2024. https://www.972mag.com/lavender-ai-israeli-army-gaza/.
- Bergengruen, Vera. "How Israel Uses AI in Gaza—and What It Might Mean for the Future of Warfare." *Time*, December 18, 2024. https://time.com/7202584/gaza-ukraine-ai-warfare/.
- Davies, Harry, Bethan McKernan, and Dan Sabbagh. "'The Machine Did It Coldly': Israel Used AI to Identify 37,000 Hamas Targets." *The Guardian*, April 3, 2024. https://www.theguardian.com/world/2024/apr/03/israel-gaza-ai-database-hamas-airstrikes.
- Davies, Harry, Bethan McKernan, and Dan Sabbagh. "'The Gospel': How Israel Uses AI to Select Bombing Targets in Gaza." *The Guardian*, December 1, 2023. https://www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets.