The Machine Cannot Stand Trial: Responsibility for War Crimes in the Age of AI

SIAI Research Editorial

Swiss Institute of Artificial Intelligence, Chaltenbodenstrasse 26, 8834 Schindellegi, Schwyz, Switzerland

Abstract

AI-enabled warfare is widely imagined as a domain in which responsibility dissolves, where a drone approaches, an algorithm classifies and the destructive act occurs so far from any human hand that no one can plausibly be blamed. This paper argues that this picture is mistaken. AI does not eliminate human agency in war; it rearranges and obscures it. Drawing on the Nuremberg principles, the contemporary practice of international humanitarian law, the documented conduct of remote and algorithmic warfare in Ukraine and Gaza and the moral psychology of distance and killing, the paper contends that accountability has to follow meaningful control, knowledge, foreseeability, legal duty and the capacity to prevent harm, not physical proximity to the victim. It develops a five-part responsibility test and applies it across the full chain of actors, from political leaders to operators to designers and procuring states. It then confronts a second, compounding danger: generative AI not only fabricates wartime content but produces a condition of ambient uncertainty in which authentic evidence can be dismissed as synthetic, yielding what has been termed a liar's dividend for perpetrators. Establishing responsibility in AI-enabled warfare therefore requires preserving two chains simultaneously, the chain of authorization and the chain of evidence. The greater danger is not that machines become guilty but that institutions use machines to make human guilt harder to locate and harder to prove.

1 Introduction - The Courtroom Without a Defendant

In the Palace of Justice at Nuremberg between 1945 and 1946, responsibility had a face and a name. The architects of aggression and atrocity sat in an identifiable row, confronted with their own signatures, their own directives, the testimony of those who had carried out their orders and the paper trail of an institution that had organized killing on an industrial scale. Whatever one thinks of the tribunal's contested legality, its central achievement was choreographic as much as jurisprudential. It placed particular human beings in a defined physical relationship to the crimes charged against them and refused to let the abstraction of the state absorb their guilt. The contrast with contemporary AI-enabled warfare is stark. A loitering drone approaches a building; a classification model scores the probability that a figure on a screen is a combatant; a decision-support system ranks a list of dwellings and recommends an order of attack; and the destructive act, when it comes, may occur with no one visibly pulling a trigger and no single moment at which a human being looked upon the victim and chose. If such a system unlawfully kills civilians, someone must still be identified to sit in the defendant's chair.

That identification is harder than it first appears and it is tempting to conclude that no one can be found. The intuition that automation creates a void of responsibility has a distinguished philosophical pedigree and it is precisely the intuition this paper sets out to reject. The conventional framing treats the problem as a novel metaphysical puzzle about machines, whether an algorithm can be blamed, whether code can bear guilt, when the more urgent problem is institutional and moral. AI distributes human agency across more hands, pushes it earlier in time, and buries it deeper inside organizations, with the practical effect that responsibility becomes harder to locate rather than actually absent. The correct analytical response traces the human decisions that gave the machine its authority: who chose to build it, who chose to trust it, under what conditions, with what knowledge and with what capacity to intervene.

A second problem compounds the first and recent work has made it unavoidable. Even where human responsibility remains traceable in principle, the evidentiary environment in which it would be proven is now under systematic assault. Generative AI injects false images and fabricated footage into wartime information flows and beyond that, it manufactures a pervasive condition of doubt in which authentic documentation can be waved away as synthetic. Research published by the Carnegie Endowment for International Peace in mid-2026 describes the risk of this doubt becoming the permanent condition of modern war documentation.[1] It produces what legal scholars have called the liar's dividend, the benefit that accrues to a wrongdoer not only from passing off fakes as real but from passing off the real as fake, since widespread awareness that fake content exists has made false claims of fake news easier to sustain.[2] When Iran's foreign minister posted an authentic photograph of rows of freshly dug graves after strikes that Iranian authorities reported had killed around 170 people, at least 110 of them schoolchildren, opposition accounts immediately labeled the image AI-generated. By the time the image was verified as genuine, the doubt had already spread and the burden fell on grieving families to prove that their dead were real.[3]

The two accountability problems are distinct but mutually reinforcing. The first concerns who is responsible when an AI-enabled system contributes to an unlawful attack. The second concerns whether that responsibility

can be established at all once the surrounding evidence can be manipulated, flooded or pre-emptively discredited. The second problem has been documented far more concretely than the first. The mechanics of evidentiary sabotage are visible in Gaza, Tigray, Sudan and Iran, while the chain of responsibility inside an algorithmic kill chain remains mainly theoretical and shielded by classification.[4] The cast of potential defendants is long. It includes the political leader who authorized the campaign and set its tolerances; the commander who selected the operational policy and the classes of target; the officer who approved a particular strike; the operator who activated the system; the intelligence analyst who accepted the machine's recommendation; the programmer who built the targeting model; the firm that sold it; and the state that procured, reviewed and deployed it. The task of this paper is not to convict any of them in the abstract but to establish the principles by which responsibility should be allocated among them, principles that, as the next section argues, were substantially settled three-quarters of a century ago and have merely been made harder to apply.

2 What Nuremberg Actually Established

Before Nuremberg, the grammar of international wrongdoing was collective. States waged war, armies committed excesses, governments pursued policies and the individuals who directed these entities could shelter behind the official character of their acts. The International Military Tribunal broke this grammar with a holding that has been quoted ever since: "Crimes against international law are committed by men, not by abstract entities and only by punishing individuals who commit such crimes can the provisions of international law be enforced."[5] The significance of this holding was not that it invented accountability for cruelty but that it relocated it, from the fictional person of the state to the actual persons who used the state as their instrument.

The principles distilled from the Charter and Judgment by the International Law Commission in 1950 are compact and remain the conceptual foundation of the field.[6] A person who commits an act constituting a crime under international law is responsible and liable to punishment. The absence of a domestic penalty does not relieve that responsibility. Acting as head of state or as a responsible government official confers no immunity. Acting under superior orders does not relieve responsibility so long as a moral choice remained genuinely possible for that person. Complicity in such a crime is itself a crime. Read together, these propositions accomplish three things at once. They substitute individual responsibility for institutional abstraction; they establish that responsibility can attach at several levels of a hierarchy rather than at only one; and they imply, through their careful conditioning, that not every death in war and not every act of a subordinate is automatically criminal. Responsibility tracks conduct, knowledge, participation, authority and the availability of a genuine choice.

It is worth dwelling on what the superior-orders provision does and does not say, because it is routinely misread. It does not abolish the relevance of orders; it denies them the status of an automatic excuse where a moral choice remained open. The standard is calibrated, not blunt. It concentrates the heaviest responsibility on those who designed and administered organized criminal violence, the architects and senior functionaries who set policy and issued directives, without thereby granting automatic absolution to those who executed them. The executioner is neither presumptively guilty nor presumptively innocent; his position on the chain of command is evidence, not verdict. Later codifications preserve this structure and sharpen it. The doctrine of

command responsibility, expressed in Article 28 of the Rome Statute, holds a military commander criminally responsible for crimes committed by forces under effective command and control where the commander knew or should reasonably have known given the circumstances at the time, that the forces were committing or about to commit such crimes and failed to take all necessary and reasonable measures to prevent or repress them.[7] This "should have known" standard is a negligence standard. A commander who receives repeated reports of unusual civilian casualties in a unit's area of operations cannot escape by cultivating ignorance.

The lesson most often drawn from Nuremberg, that guilt rises automatically to the highest office, is only half right. What the tribunal actually established was more durable and more demanding: institutions cannot make individual human responsibility disappear. Bureaucracy, hierarchy, technical specialization and the invocation of orders are precisely the ways through which responsibility is diffused until it seems to belong to no one and the entire point of the Nuremberg settlement was to deny that diffusion its exculpatory power. AI-enabled warfare now tests this same principle, because the algorithm is a new and unusually effective diffuser. If a machine can be inserted between the decision and the death, the temptation is to treat the machine as the terminus of responsibility, the point at which the human chain runs out. The remainder of this paper traces who, on a chain of algorithmically mediated killing, should still be summoned to answer.

3 Remote Warfare Can Save Soldiers While Making Killing Easier

Distance has always been the great solvent of the reluctance to kill. The military psychologist Dave Grossman argued in his study On Killing that resistance to killing declines as the distance between attacker and victim grows, from the visceral difficulty of edged-weapon combat to the near-frictionless ease of releasing ordnance from altitude.[8] Drones and remote systems extend this trajectory. They let a state see less blood and seeing less blood, they may make killing easier to authorize and easier to repeat. Any honest account of AI-enabled warfare has to hold this paradox steadily in view, because the technology's genuine benefits and its genuine dangers flow from the same source.

The benefits are real and should not be minimized in order to score a rhetorical point. Reconnaissance drones gather intelligence without exposing personnel; remotely operated systems reduce the need to send soldiers into heavily defended areas; machine vision can, in some conditions, classify objects more consistently than an exhausted or frightened operator; and a properly constrained system, bounded by tested parameters, could in principle improve precision and reduce impulsive firing. Fewer soldiers may die during reconnaissance, targeting and certain attack operations. These are not trivial considerations and a policy that ignored them in order to preserve an older and bloodier image of honorable combat would be morally unserious.

What AI-enabled warfare actually encompasses deserves precision, because the public debate tends to collapse distinct technologies into a single specter. The International Committee of the Red Cross, in its position paper on autonomous weapon systems, defines such systems as weapons that, once activated, can select and engage one or more targets without further human intervention: systems in which, after a human sets them in motion, the application of force is triggered by environmental inputs matched against a generalized target profile.[9] This is a narrow category and much of what is called AI in contemporary war falls outside it.

Machine-vision classifiers, pattern-of-life analytics and decision-support systems that recommend or select targets for human approval are, on the ICRC's own account, distinct from autonomous weapons proper. The umbrella term is useful precisely because it spans this range but the analysis must not blur the differences, since responsibility attaches differently at each point. What unites the category is the ICRC's central and, here, decisive position: it is humans, not machines, who must determine the lawfulness of an attack and that determination cannot be delegated to a machine process.[10] The legal judgment is non-transferable. A machine may perform technical tasks but the determination that a strike complies with distinction, proportionality and precaution is a context-specific human assessment for which the human remains accountable.

The following table locates the accountability concern that predominates at each level of mediation and it warrants attention because the concerns are not the same.

Table 1: **Accountability Across Levels of Weapon Mediation**

SYSTEM TYPE	TARGET IDENTIFICATION	FINAL ENGAGEMENT	PRINCIPAL ACCOUNTABILITY CONCERN
Conventional weapon	Human	Human	Orders, intent, and conduct of an identifiable actor
Remotely operated drone	Human operator	Human operator	Distance may obscure the operator and the command chain behind the screen
AI decision-support system	AI recommends or prioritizes	Human formally approves	Automation bias and superficial human review that is legally present but operationally hollow
Autonomous weapon	System selects and engages after activation	Machine process within human-set parameters	Responsibility shifts to the design, review, and authorization that preceded deployment

The movement down this table is a movement of responsibility backward in time and upward into the institution. It does not describe a leak of responsibility out of the human world; it describes its migration. This is where the paradox bites hardest. The very automation that spares one side's soldiers can, on the other side, enable more killing rather than less, because the same industrial acceleration that removes friction from the attacker's decision multiplies the number of targets that can be generated, processed and struck in a given period. Saving soldiers and allowing faster, larger-scale destruction of the adversary's people and infrastructure are not competing descriptions of the technology; they are two faces of a single capacity. What matters is not whether the capacity is used but whether the humans who wield it retain and exercise the judgment the law requires.

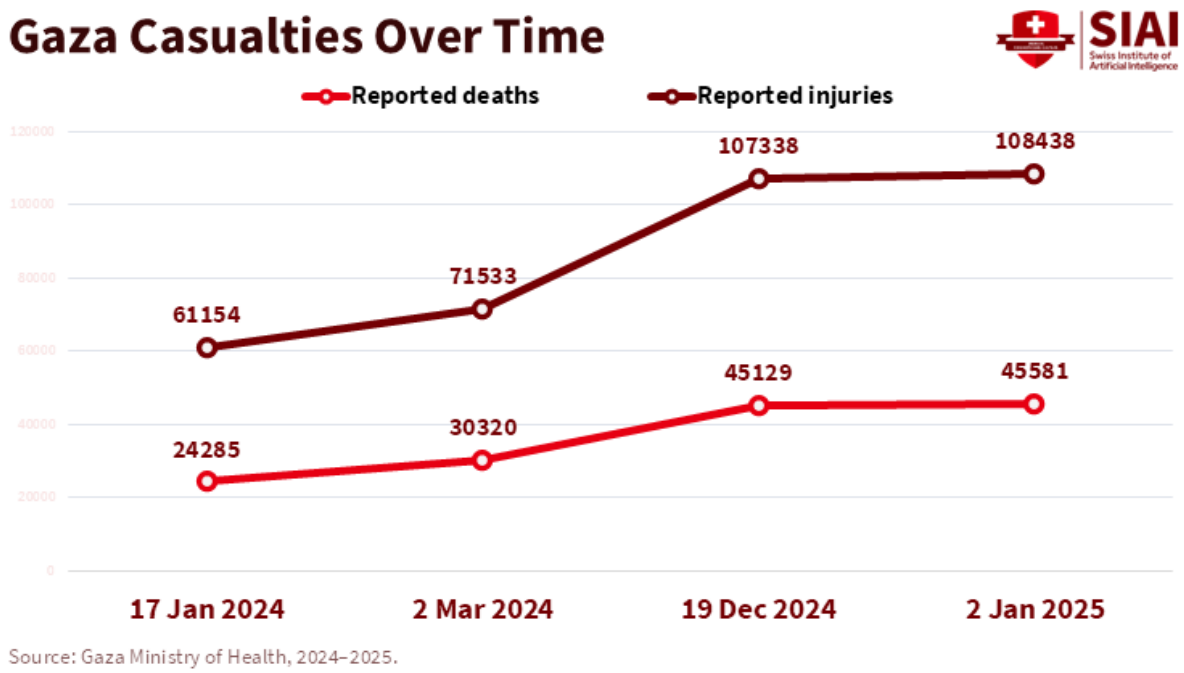

Figure 1: Deaths and injuries have moved together, not apart; the ratio between the two lines is the real story here, not the slope of either one alone.

4 Actual Cases

4.1 Ukraine: Drones Do Not Necessarily Mean Autonomous Killing

The war in Ukraine is often invoked as the proving ground of a new autonomous age but its most rigorously documented drone atrocities point in almost the opposite direction. In a conference-room paper of 28 May 2025, the UN Human Rights Council's Independent International Commission of Inquiry on Ukraine concluded that Russian armed forces had committed the crime against humanity of murder and the war crime of attacking civilians through a months-long pattern of drone strikes on the right bank of the Dnipro River in Kherson Province. The Commission documented attacks across Kherson city and sixteen other localities over a stretch of more than one hundred kilometers, in which, by the official sources it cited, nearly 150 civilians had been killed and hundreds more injured. Its findings rested on interviews with 91 residents, the geolocation of 124 of more than 300 publicly available attack videos and the review of over 600 text posts on Telegram channels.[11] The perpetrators used commercially available civilian drones weaponized to drop explosives, as well as first-person-view suicide drones, all equipped with cameras letting operators to track and aim remotely.

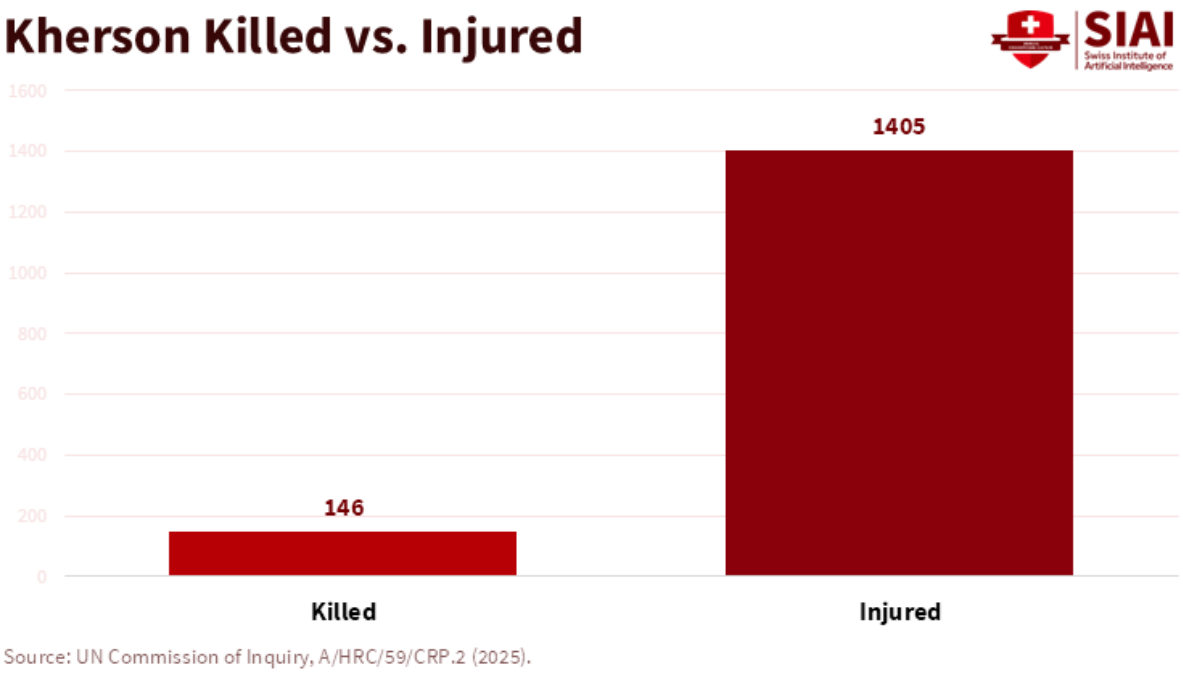

Figure 2: Roughly ten injuries for every death is a signature of sustained, deliberate targeting rather than isolated error.

The evidentiary detail that matters most for the present argument is this: the drone footage, much of it posted by the perpetrating units themselves on Telegram channels, showed that the perpetrators could clearly see the victims, leaving no doubt that they intended to target civilians.[12] Here is remote warfare in which distance created no accountability gap whatsoever. The operator saw a woman walking her dog, a hospital employee arriving at work and an ambulance and struck anyway. The command structure was identifiable, the organizational policy inferable from the pattern's scale and duration and the intent legible from the operators' own cameras. Ukraine therefore demonstrates two things at once. Technological precision does not guarantee civilian protection; a camera that lets an operator distinguish a civilian perfectly well is also a camera that lets him target one deliberately. And the remoteness of a weapon does not, by itself, dissolve responsibility. Where the human remains in the loop and the record survives, the operator and those who directed him remain as accountable as any rifleman. The accountability gap, where it exists, is manufactured, not intrinsic.

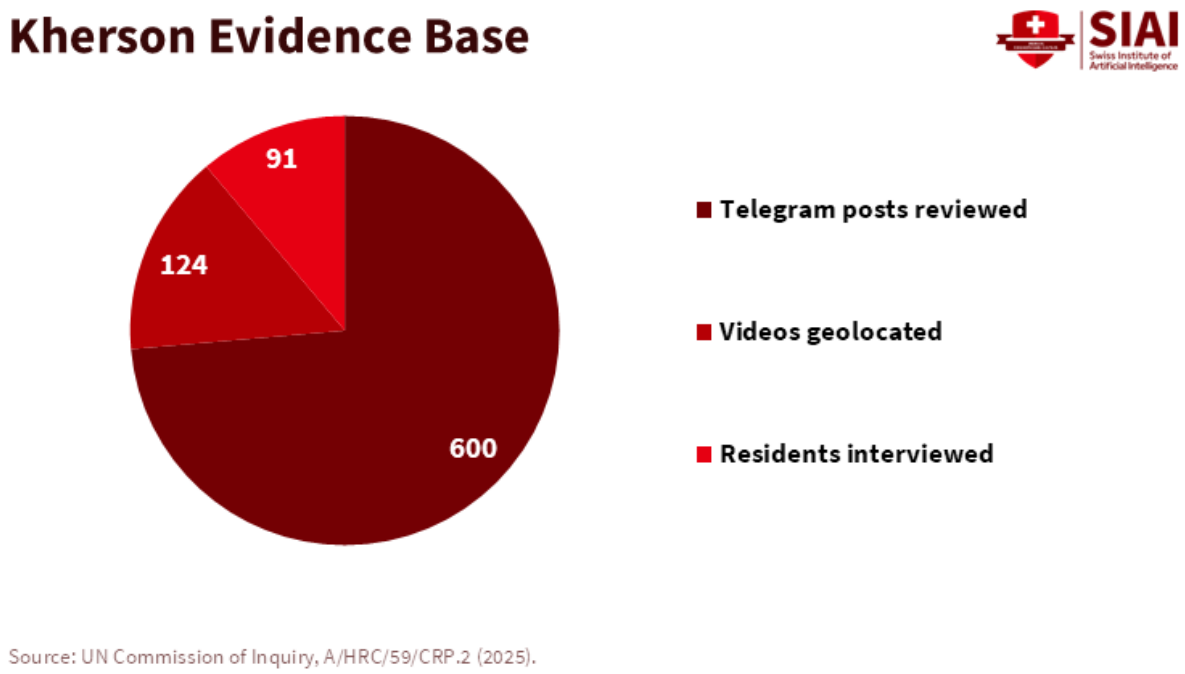

Figure 3: Almost none of it came from official disclosure, it was built from what the perpetrators posted themselves.

4.2 Gaza: AI Decision Support and the Acceleration of Targeting

Gaza presents the harder case, because there the reported role of AI sits upstream of the trigger, in the generation and ranking of targets rather than the firing of weapons. An investigation published by the Israeli-Palestinian outlet +972 Magazine and Local Call in April 2024, based on the testimony of six Israeli intelligence officers, reported that the Israeli military used a system named Lavender that identified as many as 37,000 Palestinians and their homes as suspected militants and possible strike targets; a related system reported earlier, The Gospel, generated building and infrastructure targets. Sources told the outlet that personnel devoted roughly twenty seconds to each target before authorizing a bombing, often merely confirming that the marked individual was male, despite knowing that the system's classifications were erroneous in an estimated ten percent of cases.[13] The Israeli military rejected the reporting, its spokesperson describing the tools as auxiliary aids that assist officers and require independent analyst confirmation and stating that the systems do not autonomously select targets for attack.[14]

These claims must be characterized with care, because they are contested and their legal weight depends on their accuracy. The reporting rests on anonymous testimony and has not been adjudicated. Independent UN human rights experts, however, took it seriously enough to state in April 2024 that the reported use of systems such as Gospel, Lavender and Where's Daddy, combined with lowered human due diligence, would, if accurate, help explain the extraordinary toll in Gaza, noting that more than 15,000 deaths occurred in the first six weeks, when AI systems appear to have been most heavily relied upon.[15] The responsible analytical posture is to treat the specific mechanics as reported and alleged rather than established, while recognizing that even the uncontested contours, that AI systems accelerated target generation and that humans retained formal although compressed review- are sufficient to frame the accountability question. What must be resisted is the

slide from "AI helped generate targets at scale" to "the machine decided," because that slide is exactly the exculpatory move this paper exists to refuse. As the journalist behind the investigation stressed, the decision to bomb private homes systematically was made by humans; the AI let a deadly practice be multiplied at scale.[16]

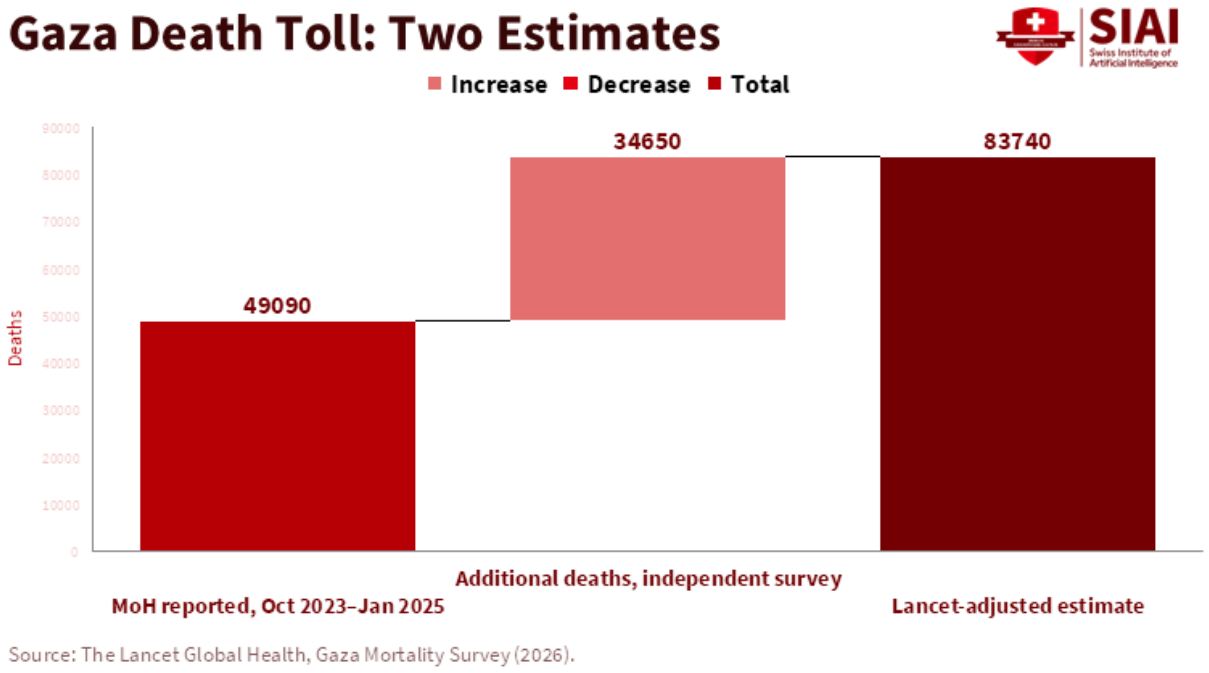

Figure 4: The gap between the two bars is larger than the entire original count.

4.3 Distance, Scale and Moral Inhibition

Beneath the legal argument runs a moral-psychological one that the Gaza and Ukraine cases jointly illuminate. Traditional frontline violence forces the soldier into the human reality of killing, with all the resistance and residue that Grossman documented. Remote warfare removes the attacker from physical danger and AI-assisted targeting can go further, removing the decision-maker from psychological confrontation with the individual victim altogether. When a target arrives as a name on a ranked list with a probability value and a narrow window for approval, the deliberative encounter that once attended the choice to kill is replaced by an act of validation. The more targets a system produces, the less time any human spends examining each one. Careful judgment degrades into industrial processing and the twenty-second review becomes not an aberration but the predictable equilibrium of a system optimized for tempo.

Yet the moral psychology is more ambivalent than a simple thesis of moral disengagement would allow and honesty requires accepting the complication. Interviews with British Reaper crews conducted by Peter Lee show that physical distance does not reliably produce emotional distance.[17] Documented rates of post-traumatic stress and what researchers term moral injury among remote drone operators point the same way.[18] The high-resolution camera that enables killing from half a world away also compels the operator to watch the blast, the survivors, the bodies gathered into blankets and to return to the same screen the next day. The relevant distinction is not physical but epistemic and moral distance, which surveillance technology can paradoxically narrow even as it extends physical range. This ambivalence does not rescue the technology;

it sharpens the argument. The danger of AI-enabled targeting is not that every operator becomes a callous button-pusher but that the institution can be designed to minimize the moments of confrontation that generate moral friction and to distribute the decision so widely that no individual experiences himself as the author of the death. Technology can reduce one kind of brutality, the sacrifice of soldiers, while enabling another: more frequent, faster and emotionally insulated applications of force. The absence of blood on the operator's hands does not mean no human decision occurred; more likely, the decision was made earlier, farther away and inside a more complex institution.

5 Who Should Be Blamed When an Algorithm Contributes to an Unlawful Killing?

A machine cannot be a moral defendant. The philosophical literature that first raised the responsibility gap, most influentially Robert Sparrow's 2007 argument that no one can be fairly held responsible for the wartime acts of a genuinely autonomous system, since neither the programmer nor the commander could predict its behavior and the machine itself cannot be punished, identified a real puzzle but drew from it a conclusion that does not follow.[19] Sparrow's own remedy was that systems like that should therefore not be deployed. But the systems in use today are not the fully autonomous agents of the thought experiment. They are decision-support tools and human-supervised weapons whose behavior, while imperfectly predictable, is foreseeable in exactly the way the law requires. Later philosophers have argued that a commander who activates a system knowing it will operate autonomously and knowing that error isn't simple possible but statistically certain satisfies both the control condition and the foreseeability condition for moral responsibility.[20] The gap closes because the human who empowered the machine never ceased to be responsible for it. The question is which humans decided to give the system authority, under what conditions, with what knowledge, with what safeguards, not whether the algorithm is guilty.

5.1 Responsibility Test

Rank and job title alone should never determine conviction, because they are proxies for the things that actually ground responsibility rather than the things themselves. Five factors do the real work. The first is causal contribution: how directly and substantially the person's act or omission contributed to the unlawful attack. The second is knowledge and foreseeability: what the person actually knew and what a reasonable person in that role should have anticipated. The third is meaningful control: whether the person could authorize, halt, modify, or prevent the operation, a concept first proposed by the non-governmental organization Article 36 in 2013 to describe control over a weapon's critical functions of selecting and engaging targets, which subsequently migrated into the wider ethics of AI.[21] The fourth is legal and institutional duty: whether the person was charged with reviewing, supervising, or constraining the system. The fifth is intent and acceptance of risk: whether the person intended the unlawful outcome or knowingly accepted an unlawful risk. These factors are cumulative and weighted rather than a checklist to be tallied. A low causal contribution paired with high

knowledge and a clear duty may ground more responsibility than a high causal contribution undertaken in excusable ignorance.

Political and military leadership occupy the position where the highest-command intuition is strongest and most defensible. Leaders who establish unlawful targeting policies, who authorize indiscriminate or foreseeably disproportionate operations, who remove or hollow out meaningful human review in the name of tempo, who ignore recurring and reported civilian harm or who fail to investigate and discipline known violations bear responsibility that no downstream automation can discharge. If a policy sets an acceptable civilian toll per target category, the wrongfulness of the resulting deaths originates in that policy, not in the model that populated the target list. The model executed a tolerance that human leaders chose. This is the precise sense in which those who construct the operational system cannot escape by pointing to the final operator or the machine.

Operational commanders and strike authorizers usually stand closest to the legally decisive deployment decision and the ICRC's analysis of autonomous systems maps their duties with unusual precision. Because the user of such a system cannot know in advance which specific person or object will trigger a strike, the commander must assess lawfulness prospectively, at the moment of activation, across the entire area and duration of the system's operation and in light of all reasonably foreseeable changes in circumstances.[22] This translates into a demanding set of responsibilities: selecting the environment in which the system is used, restricting the classes of target, setting geographic and temporal limits, assessing proportionality and precautions, judging whether the system is sufficiently predictable to be used at all and canceling or suspending the attack when circumstances change.

Operators and intelligence analysts should not be presumed guilty, and they should not be presumed innocent either and here the concept of the moral crumple zone is indispensable. The anthropologist Madeleine Elish uses the term to describe how, in complex automated systems, responsibility for a failure is misattributed to the human operator who in fact had limited control over the system's behavior. The human becomes a component that absorbs the moral and legal responsibility when the overall system malfunctions, while the integrity of the technical system and of those who designed and deployed it, is protected.[23] A twenty-second reviewer, given no real authority or time to reject the machine's output, is a candidate crumple zone. Blaming him alone would replicate the very injustice the concept warns against. The relevant considerations are specific: whether the analyst knowingly approved a civilian target, whether reliable contrary evidence sat before him, whether he was permitted sufficient time and genuine authority to reject the system's recommendation or whether rejection was institutionally discouraged and whether he intentionally disregarded visible signs of civilian presence. Automation bias, the well-documented human habit of over-trusting and under-scrutinizing automated recommendations, can reduce a nominal human decision to a rubber stamp while the paperwork still records a human choice. It is a genuine institutional and psychological phenomenon and it cuts in two directions.[24] It mitigates the individual operator's culpability where the institution engineered the conditions for it and it aggravates the culpability of those who designed those conditions knowing what they would produce.

Designers and manufacturers require the most careful distinctions, because the intuitions here are the least reliable. Building a targeting algorithm is not, in itself, the commission of a war crime, any more than designing a missile makes an engineer responsible for every unlawful launch. The default position must be that the

programmer is neither the presumptive villain nor categorically immune. Potential responsibility arises, however and international criminal law's doctrines of complicity and aiding and abetting can reach it, where a designer or firm builds a system specifically to facilitate unlawful targeting, knowingly conceals dangerous defects, falsifies accuracy or safety results, continues to provide essential assistance while knowing the system is being used for criminal operations or intentionally strips out the safeguards required for lawful use. The precedents are sparse but real. Post-war tribunals convicted industrialists and national courts have convicted businessmen such as the Dutch traders Frans van Anraat and Guus Kouwenhoven for aiding and abetting atrocities through the supply of means.[25] The threshold is demanding. The Rome Statute limits the Court to natural persons and its aiding-and-abetting standard requires that assistance be rendered specifically to facilitate the crime, a stringent mental-state requirement that a profit-motivated firm can often plausibly deny but it is not empty.[26]

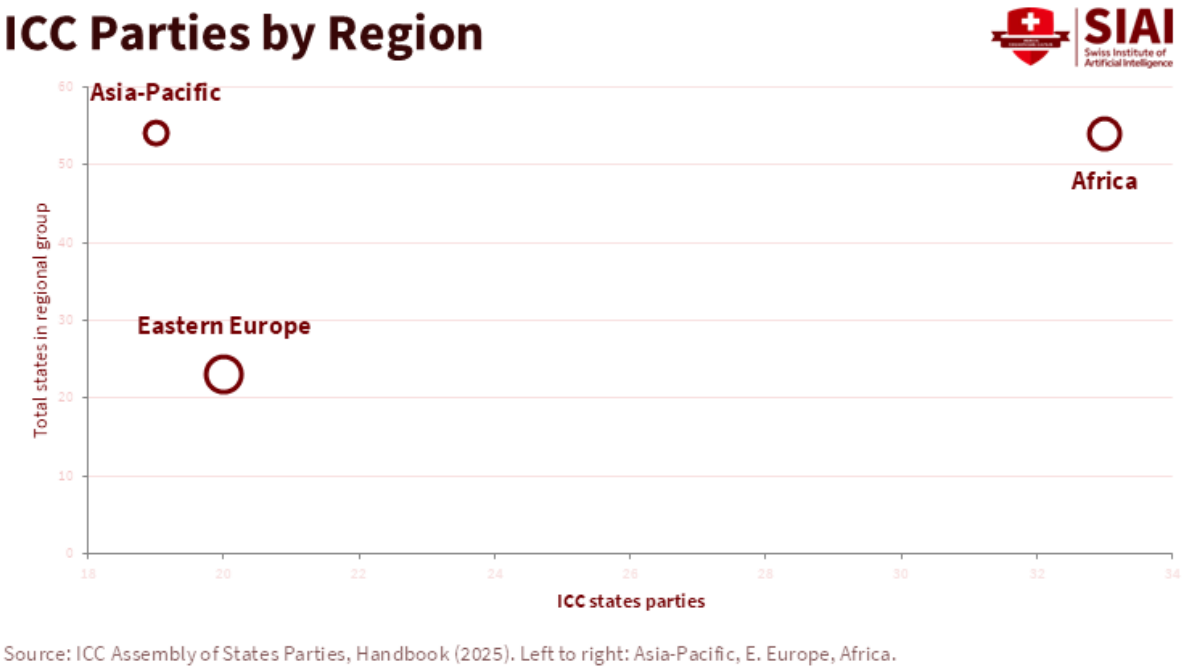

Figure 5: Coverage tracks geopolitics more than geography: near-universal in Eastern Europe, barely a third of Asia-Pacific.

States, procurement bodies and reviewing institutions bear a responsibility that precedes the battlefield entirely and it is codified. Article 36 of Additional Protocol I to the Geneva Conventions provides that "in the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited" by international law.[27] This obligation is broad. It reaches acquisition, not only development, so a purchasing state cannot outsource the judgment to the manufacturer's home country and it is widely regarded as binding on all states regardless of treaty ratification, because it follows reasonably from the prohibition on using unlawful weapons. Article 36 thus creates institutional responsibility upstream of any operator's decision. A state that fields an AI-enabled system whose effects it cannot understand, predict and explain has failed a legal duty before a single strike occurs. The concrete demands that follow from taking this seriously are specific and assignable: pre-deployment legal review of the kind Article 36 requires, tested and documented operational

limits, traceable command authorization for each deployment, preservation of system logs, documented override and cancellation procedures and independent post-strike investigation.

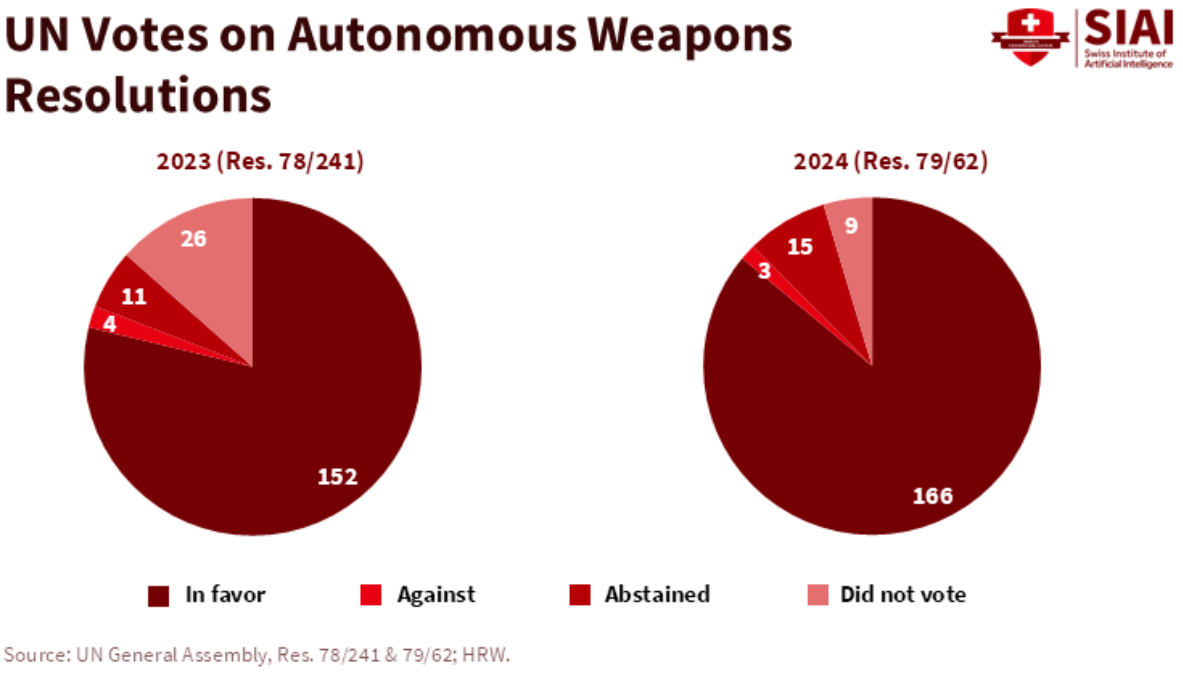

Figure 6: Opposition didn't grow between 2023 and 2024 — it shrank, even as the technology matured.

5.2 Generative AI and Atrocity Denial

Suppose, however, that every link in this chain of authorization could in principle be traced. A second technology now works to ensure that it cannot be proven. The generative AI that produces synthetic images and video adds falsehoods to the wartime record, and, worse, degrades the evidentiary value of the entire record; it degrades the evidentiary value of the entire record, because once convincing fakes are known to exist, every authentic document becomes vulnerable to the accusation that it, too, is fabricated. This is the liar's dividend described earlier: bad actors benefit not only from passing fakes off as real but from passing real evidence off as fake.[28] The mechanism is now visible across theatres. In Gaza, an authentic photograph of an injured man at Al-Shifa Hospital was doctored with an AI-added third leg and then circulated as supposed proof that images of Palestinian victims are generally AI-generated; a displaced man raising money online was accused of being an AI creation on the strength of garbled text on a mass-produced blanket and spent hours proving he was real.[29] In Iran, a separate AI-generated image of blood-soaked streets, posted by a partisan commentator and swiftly debunked, nonetheless served the regime's interest by lending credence to the claim that all documentation of a real massacre was suspect.[30] In Tigray, communications blackouts achieved the older version of the same end, making documentation dangerous or impossible, while AI now achieves it in a decentralized, industrial form available to anyone with a model and an account.[31]

The critical insight is structural: an AI-generated lie need not survive verification to succeed. The accusation does not need to persuade; it only needs to plant enough doubt that the truth arrives already damaged. Sometimes the corruption runs the other way, with authentic footage paired with a deceptive frame, as when

real video from a militarized aid site was deployed as propaganda and then, when accused of being synthetic, defended by forensic analysts who found no sign of AI even as the political framing around it remained a lie. The accusation of fakery has, in this environment, no fixed target. It is a weapon available to every party and it escapes the control of those who first reach for it. The harm, meanwhile, is concentrated on specific and identifiable people: the parents burying children whose graves are called synthetic, the survivor forced to prove his own existence, the witness whose evidence is discredited before it is even examined.

The implication for accountability is that AI-era responsibility must preserve two chains at once and neither suffices alone. The first is the chain of authorization: who designed, approved, procured, deployed, supervised and activated the system, recorded in a form that can survive institutional denial. Without it, responsibility becomes diffuse to the point of vanishing, exactly as the machine's defenders intend. The second is the chain of evidence: how targeting data, system outputs, human approvals, strike logs, images and testimony were recorded and preserved. Without it, responsibility may exist in principle but cannot be proven in practice. The two chains fail in complementary ways. An institution can defeat accountability either by diffusing the decision until no author remains or by degrading the record until no proof survives. This is why practical remedies now emerging from the human-rights and open-source-investigation communities, provenance and content-credential infrastructure embedded at the point of capture, distributed and tamper-resistant archives, chain-of-custody standards such as the Berkeley Protocol on Digital Open Source Investigations and protections for the physical safety of those who document, are not peripheral technical housekeeping but the evidentiary counterpart of the Nuremberg principle itself.[32]

6 Technology Changes the Weapon, Not the Need for Judgment

Return, then, to the courtroom. Nuremberg's lasting lesson was never merely that senior leaders can be punished. It was that organized institutions cannot transform human crimes into ownerless events, that layers of hierarchy and technical specialization cannot turn a superior order into an excuse that erases any single author. AI-enabled warfare revives this temptation in a new and more sophisticated form. Its logic surfaces in the anticipated chorus of mutual deferral: the commander blames the model, the operator blames the recommendation, the programmer blames the user, the manufacturer blames the state, the political leader blames an operational failure and the state, reaching for the newest instrument of all, dismisses the very evidence as synthetic. Each move is individually plausible. Together they amount to a mechanism for manufacturing impunity, the same mechanism Nuremberg was built to dismantle, now re-tooled.

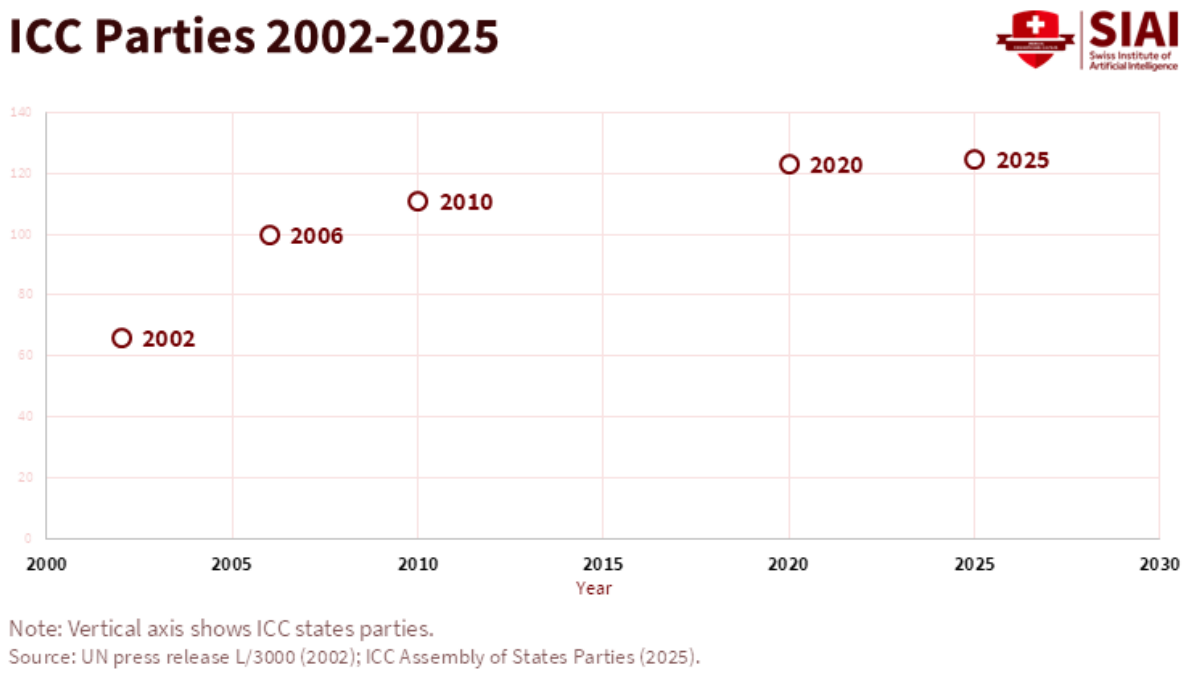

Figure 7: Almost all the growth happened in the Court's first decade; the last fifteen years added almost nothing.

The counterarguments deserve better than dismissal, because at their strongest they are correct as far as they go. Remote and AI-enabled weapons can indeed spare soldiers who would otherwise be sacrificed and it would be morally simplistic, even perverse, to demand that human beings be exposed to danger merely to preserve an older and more heroic image of war. Automation can, under the right constraints, reduce certain impulsive human errors. And it is true that warfare has always been technologically mediated, that the drone is in one sense only the latest point on a continuum that runs from the sling to the artillery piece. These arguments establish that AI-enabled warfare is not uniquely evil and need not be categorically prohibited. What they do not establish is that removing people from the frontline may also remove human decision-making from the decision to kill. The continuum of mediation is real but it has always been accompanied by a continuum of responsibility that tracks control, knowledge and duty rather than proximity. The artillery officer beyond sight of his target was never thereby absolved. The novelty of AI is not that it mediates but that it can be made to obscure and the appropriate response is to insist that the obscuring be undone: that logs be kept, reviews be genuine, authorizations be traceable and evidence be preserved.

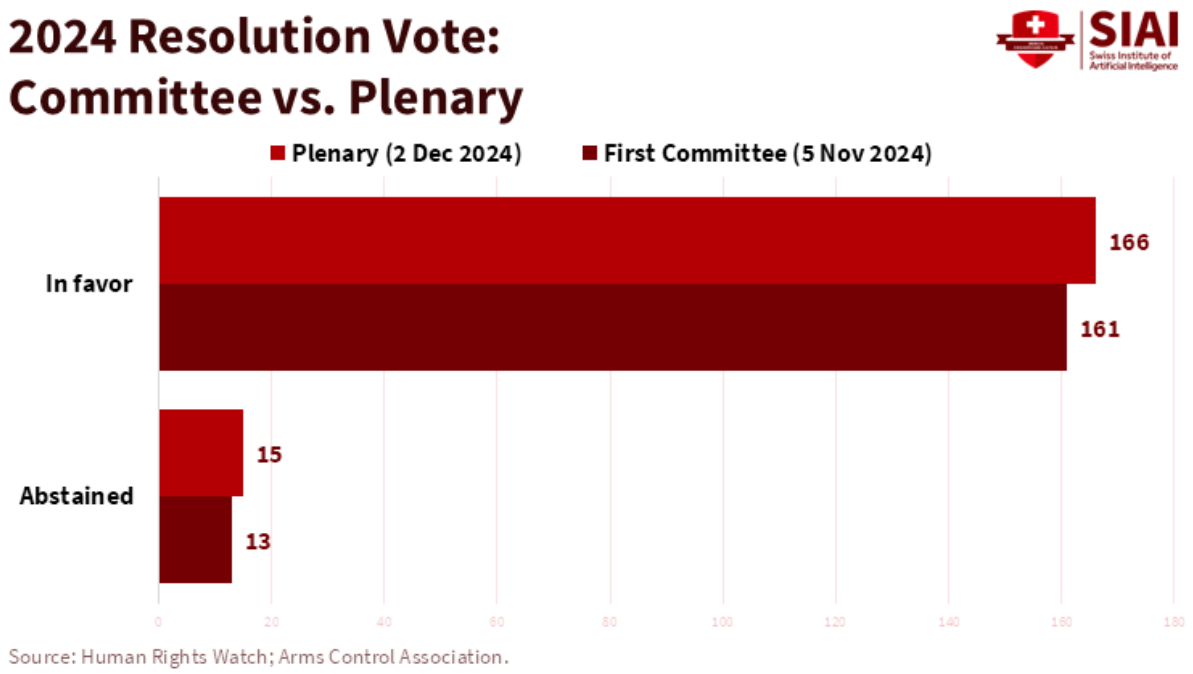

Figure 8: Support hardened rather than softened between the two votes; the opposite of what backroom pressure usually produces.

The weapon has changed. Responsibility has not vanished. A machine may identify a target and even trigger the attack but human institutions decided to build it, chose to authorize it and placed living people within its reach. Those decisions are where the moral weight resides and they are made by people with names. The machine cannot stand trial. The people who gave it that power still can.

7 Conclusion - Accountability for AI-Enabled War Crimes

The reframing this paper has urged is simple but consequential. AI-enabled warfare does not create a void of responsibility; it creates a diffusion of responsibility and the two are easily confused because they look alike from the outside. The apparent void is an artifact of institutional design, of decisions distributed across many hands, pushed backward in time, and buried inside organizations. It can be undone by the same means Nuremberg used against the abstraction of the state: by insisting that crimes are committed by people, and by tracing the chain of human decisions that gave the machine its authority. Responsibility should follow meaningful control, knowledge, foreseeability, legal duty, and the capacity to prevent harm, applied without sentiment across leaders, commanders, operators, designers, and states, and calibrated by the recognition that the nearest human operator is often the least culpable, not the most.

The evidence assembled here supports a precise policy conclusion rather than a general exhortation and the international machinery to act on it already exists in embryo. In October 2023 the UN Secretary-General and the President of the ICRC jointly called on states to conclude negotiations on a legally binding instrument on autonomous weapons by 2026.In December 2024, the General Assembly adopted Resolution 79/62 by 166 votes to 3, with only Belarus, North Korea and Russia opposed. Since January 2023, United States doctrine has required that certain systems allow commanders and operators to exercise appropriate levels of human

reasoning over the use of force. The task is to give these commitments teeth. Because responsibility in AI-enabled warfare fails in two distinct ways, through the diffusion of the decision and through the degradation of the proof, the governing requirement must be that both chains be preserved by law and by design: a chain of authorization traceable to named individuals at every step and a chain of evidence preserved against both fabrication and denial. Concretely, no AI-enabled system whose effects cannot be understood, predicted and explained should pass an Article 36 review; human review must be resourced with the time and authority to be genuine rather than a rubber stamp; system logs, override records and post-strike investigations must be mandatory and independently auditable; and provenance infrastructure and protected archives must be treated as instruments of accountability. The machine cannot be blamed and it cannot be believed or disbelieved on its own authority. Only people can be and the work of law and policy is to ensure that when an algorithm kills unlawfully, the people who gave it that power can still be found, named and judged.

References

[1, 3, 4, 29, 30, 31] Alimardani, M. and Rigot, A. (2026) *Generative AI and Atrocity Denial in War.* Washington, DC: Carnegie Endowment for International Peace.

[2, 28] Chesney, R. and Citron, D. (2018) 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security', *California Law Review*, 107, pp. 1753–1819.

[5] International Military Tribunal (1946) *Judgment of the International Military Tribunal for the Trial of German Major War Criminals.* Nuremberg: IMT.

[6] International Law Commission (1950) *Principles of International Law Recognized in the Charter of the Nürnberg Tribunal and in the Judgment of the Tribunal.* New York: United Nations.

[7, 26] United Nations (1998) *Rome Statute of the International Criminal Court.* The Hague: International Criminal Court.

[8] Grossman, D. (2009) *On Killing: The Psychological Cost of Learning to Kill in War and Society.* Rev. ed. New York: Back Bay Books.

[9, 10, 22] International Committee of the Red Cross (2025) *Autonomous Weapon Systems and International Humanitarian Law: Selected Issues.* Geneva: ICRC.

[11, 12] United Nations Human Rights Council, Independent International Commission of Inquiry on Ukraine (2025) "They Are Hunting Us": *Systematic Drone Attacks Targeting Civilians in Kherson.* A/HRC/59/CRP.2. Geneva: OHCHR.

[13, 14, 16] Abraham, Y. (2024) 'Lavender: The AI Machine Directing Israel's Bombing Spree in Gaza', *+972 Magazine*, April.

[15] Office of the United Nations High Commissioner for Human Rights (2024) *Gaza: UN Experts Deplore Use*

of Purported AI to Commit 'Domicide' in Gaza. Geneva: OHCHR.

[17, 18] Lee, P. (2018) *Reaper Force: The Inside Story of Britain's Drone Wars*. London: John Blake.

[19] Sparrow, R. (2007) 'Killer Robots', *Journal of Applied Philosophy*, 24(1), pp. 62–77.

[20] Himmelreich, J. (2019) 'Responsibility for Killer Robots', *Ethical Theory and Moral Practice*, 22(3), pp. 731–747.

[21] Article 36 (2013) *Killer Robots: UK Government Policy on Fully Autonomous Weapons*. London: Article 36.

[23, 24] Elish, M.C. (2019) 'Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction', *Engaging Science, Technology, and Society*, 5, pp. 40–60.

[25] Bryk, L. and Saage-Maaß, M. (2019) 'Individual Criminal Liability for Arms Exports under the ICC Statute', *Journal of International Criminal Justice*, 17(5), pp. 1117–1137.

[27] Diplomatic Conference on the Reaffirmation and Development of International Humanitarian Law (1977) *Protocol Additional to the Geneva Conventions of 12 August 1949 (Protocol I)*. Geneva: ICRC.

[32] Human Rights Center, University of California, Berkeley and Office of the United Nations High Commissioner for Human Rights (2022) *Berkeley Protocol on Digital Open Source Investigations*. Berkeley: UC Berkeley.