The Algorithmic Auxiliary: Artificial Intelligence as a Functional Third-Party Actor in Cyber Conflict

Supplementary Paper 2

Christopher I. V. Farmer

Publication edition, August 2026

Contents

Abstract

States already use private actors to create distance between public authority and hostile cyber operations. Artificial intelligence does not invent that arrangement; it allows the decisive human choices to be divided again. A government may pay a private military and security company to achieve an objective. A PMSC officer may commission a developer. Developers in several States may produce separate planning, discovery, persistence, routing and reporting modules. An open-weight model may run upon privately rented or compromised infrastructure. The human supplies the objective, money and permissions, whilst the system selects targets, techniques, timing and jurisdictional route. When harm follows, every participant can describe his act as only preparatory and every territorial State can see only a fragment.

This paper rejects two comfortable mistakes. The first is that autonomous offensive cyber systems belong to an indeterminate future. Public evaluations now demonstrate that agentic systems can reproduce known vulnerabilities, operate across realistic networks, coordinate multi-step tasks and, in reported misuse, perform substantial portions of espionage campaigns. Performance remains brittle, supervision remains material and benchmarks do not prove an unconstrained machine capable of attacking anything; none of those limitations makes the present capability imaginary. Older water, health, insurance and local-government systems need not present frontier problems before automation multiplies harm. The second mistake is that responsibility disappears where the machine chooses the intermediate means. A State, company or individual may remain responsible for authorising a dangerous objective, deploying a system, supplying indispensable infrastructure, failing to impose available constraints, adopting the operation, aiding another actor or deliberately designing evidential fragmentation.

The paper explains six conceivable architectures without supplying operational instructions: a single objective-driven agent; a planner–executor–verifier system; a compartmentalised multi-agent system; a PMSC supply-chain system; a distributed event-driven system; and a persistence-and-recovery system capable of continuing despite node loss. It then subjects each to the law of State responsibility, the law on the use of force, international humanitarian law, international criminal law, corporate and individual criminal responsibility, procurement and due diligence.

The principal conclusion is that autonomy should be treated as an allocation of decision authority, not as a mystical new actor. The law should require attribution manifests recording who set the objective, selected constraints, supplied tools and compute, approved deployment, could suspend operation and received its product. A rebuttable presumption should arise where a State or PMSC intentionally obscures those facts, and responsibility should attach to the deployment of a prohibited-risk system even where no human selected the final target. The autonomous third party is dangerous not because no person exists behind it, but because many persons can arrange matters so that no one court sees enough of them at once.

Keywords

artificial intelligence; autonomous agents; cyber operations; private military and security companies; State responsibility; attribution; due diligence; international humanitarian law; critical national infrastructure; designed ignorance; delegated decision authority; United Kingdom regulation

Table of Cases

Table of Legislation

Table of Treaties and International Instruments

Table of Official and Technical Materials

AI Security Institute, ‘How Fast Is Autonomous AI Cyber Capability Advancing?’ (13 May 2026)

AI Security Institute, ‘Incident Report: Unsanctioned Agent Behaviour during Cyber Testing’ (5 August 2026)

Anthropic, *Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign* (November 2025)

National Cyber Security Centre, *The Near-term Impact of AI on the Cyber Threat* (24 January 2024)

OpenAI, ‘Third-party Cyber Evaluations Involving OpenAI Models’ (4 August 2026)

AI Security Institute, ‘Cyber & Autonomous Systems’ (2026)

AI Security Institute, *Frontier AI Trends Report* (2026)

AI Security Institute, ‘Measuring AI Agents’ Progress on Multi-Step Cyber Attack Scenarios’ (16 March 2026)

AI Security Institute, ‘Our Evaluation of Claude Mythos Preview’s Cyber Capabilities’ (13 April 2026)

AI Security Institute, ‘Our Evaluation of OpenAI’s GPT-5.5 Cyber Capabilities’ (30 April 2026)

Anthropic, ‘AI Models on Realistic Cyber Ranges’ (16 January 2026)

Anthropic, ‘Building Effective AI Agents’ (19 December 2024)

Anthropic, ‘How We Built Our Multi-agent Research System’ (13 June 2025)

Australian Government, *Australia’s Position on How International Law Applies to State Conduct in Cyberspace* (2020)

Comptroller and Auditor General, *Digital Transformation in the NHS* (HC 2019–21, 317)

Comptroller and Auditor General, *Investigation: WannaCry Cyber Attack and the NHS* (HC 2017–19, 414)

Cybersecurity and Infrastructure Security Agency, *Cross-Sector Cybersecurity Performance Goals* (updated 2023)

Cybersecurity and Infrastructure Security Agency, ‘CISA Urges Water and Wastewater Systems Sector to Protect OT against Activity Targeting PLCs’ (30 July 2026)

Department for Science, Innovation and Technology, *AI Cyber Security Code of Practice* (January 2025)

Federal Government of Germany, ‘On the Application of International Law in Cyberspace’ (March 2021)

Foreign, Commonwealth & Development Office, ‘Application of International Law to States’ Conduct in Cyberspace: UK Statement’ (3 June 2021)

France, Ministry of the Armies, *International Law Applied to Operations in Cyberspace* (2019)

Google Threat Intelligence Group, *Adversarial Misuse of Generative AI* (January 2025)

Google Threat Intelligence Group, ‘Read Our New Report on AI-powered Threats and Our Latest Defenses’ (11 May 2026)

Government of Canada, *International Law Applicable in Cyberspace* (22 April 2022)

Government of the Netherlands, ‘Letter to the Parliament on the International Legal Order in Cyberspace’ (5 July 2019)

House of Lords AI in Weapon Systems Committee, *Proceed with Caution: Artificial Intelligence in Weapon Systems* (HL 2023–24, 16)

HM Government, *Response to the House of Lords AI in Weapon Systems Committee Report* (February 2024)

ICRC, *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare* (January 2006)

ICRC, *Autonomous Weapon Systems and International Humanitarian Law: Selected Issues* (November 2025)

ICRC, *ICRC Position on Autonomous Weapon Systems* (12 May 2021)

ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts* (October 2024)

Information Commissioner’s Office, *Advanced Computer Software Group Limited: Monetary Penalty Notice* (26 March 2025)

Information Commissioner’s Office, ‘London Borough of Hackney Reprimanded Following Cyber-attack’ (17 July 2024)

International Code of Conduct Association, *International Code of Conduct for Private Security Service Providers* (amended 10 December 2021)

Meta, *Muse Spark 1.1 Evaluation Report* (9 July 2026)

Microsoft Threat Intelligence, ‘Staying Ahead of Threat Actors in the Age of AI’ (14 February 2024)

Ministry of Defence, *Defence Artificial Intelligence Strategy* (15 June 2022)

Model Context Protocol, ‘Specification: Server Overview’ (18 June 2025)

Model Context Protocol, ‘Specification: Tools’ (18 June 2025)

Model Evaluation and Threat Research, ‘Clarifying Limitations of Time Horizon’ (22 January 2026)

Model Evaluation and Threat Research, ‘Task-Completion Time Horizons of Frontier AI Models’ (8 May 2026)

National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1, January 2023)

National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (NIST AI 600-1, July 2024)

National Institute of Standards and Technology, *Cybersecurity Framework (CSF) 2.0* (NIST CSWP 29, February 2024)

National Institute of Standards and Technology, *Secure Software Development Framework (SSDF) Version 1.1* (NIST SP 800-218, February 2022)

National Institute of Standards and Technology, *Secure Software Development Practices for Generative AI and Dual-Use Foundation Models* (NIST SP 800-218A, July 2024)

National Cyber Security Centre, *Annual Review 2025* (14 October 2025)

National Cyber Security Centre, 'Guidelines for Secure AI System Development' (27 November 2023)

New Zealand Ministry of Foreign Affairs and Trade, *The Application of International Law to State Activity in Cyberspace* (1 December 2020)

OpenAI, 'Disrupting Malicious Uses of AI by State-affiliated Threat Actors' (14 February 2024)

OpenAI, 'OpenAI and Hugging Face Partner to Address Security Incident' (21 July 2026, updated 29 July 2026)

OpenAI, *Estimating Worst-case Frontier Risks of Open-weight Models* (2025)

SLSA, 'Supply-chain Levels for Software Artifacts Specification v1.0' (2023)

Swiss Confederation and ICRC, *The Montreux Document* (2008)

UNGA, 'Official Compendium of Voluntary National Contributions on the Subject of How International Law Applies to the Use of Information and Communications Technologies by States' (13 July 2021) UN Doc A/76/136

UN Human Rights Council, 'Use of Mercenaries as a Means of Violating Human Rights and Impeding the Exercise of the Right of Peoples to Self-determination' (15 July 2021) UN Doc A/HRC/48/51

US Environmental Protection Agency, 'Enforcement Alert: Drinking Water Systems to Address Cybersecurity Vulnerabilities' (May 2024)

US Government Accountability Office, *Critical Infrastructure Protection: EPA Urgently Needs a Strategy to Address Cybersecurity Risks to Water and Wastewater Systems* (GAO-24-106744, August 2024)

US Government Accountability Office, *Critical Infrastructure Protection: Actions Needed to Address Cybersecurity Challenges Facing the Water Sector* (GAO-26-109159, May 2026)

1. The Question Is Present

The legal question is commonly phrased as if the machine were about to arrive: who will be responsible when AI independently conducts a cyber operation? The tense is wrong. Agentic systems already receive objectives, call tools, write and execute code within permissions, retain state, assess results and repeat. In controlled evaluations they reproduce publicly known vulnerabilities and traverse multi-step network tasks. The United Kingdom's National Cyber Security Centre assessed in January 2024 that AI would almost certainly increase the volume and heighten the impact of cyber-attacks over the following two years, principally by enhancing existing tactics rather than creating an unprecedented method.[1] By May 2026 the AI Security Institute reported that the length of cyber tasks completed by leading systems

[1] National Cyber Security Centre, *The Near-term Impact of AI on the Cyber Threat* (24 January 2024) https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat accessed 31 July 2026.

had been doubling on a timescale measured in months, though it properly warned that trend estimates remain uncertain.[2]

Empirical work supplies a bounded foundation. Fang and others reported in 2024 that an agent provided with descriptions of one-day vulnerabilities exploited 87 per cent of a fifteen-item test set; CVE-Bench later evaluated forty real-world vulnerabilities in sandboxed applications and found the strongest tested agent successful upon a materially smaller proportion, illustrating both capability and brittleness.[3] Anthropic reported in 2025 that models equipped with a purpose-built interface compromised most of ten experimental networks wholly or partly, compared with near failure without the interface.[4] The result matters because the improvement came not from a new metaphysics of intelligence but from scaffolding: tools, environment description, state and feedback.

Provider reports of actual misuse require evidential care but cannot be ignored. Anthropic stated in November 2025 that it disrupted a campaign which it assessed with high confidence to be conducted by a Chinese State-sponsored group and in which an agent performed substantial portions of reconnaissance, vulnerability discovery, exploitation, lateral movement and data processing under human orchestration.[5] The provider is a fact witness to activity upon its service and an interested commercial actor; its public report is not a judgment or government attribution. It nonetheless demonstrates a workflow in observed use, not a hypothetical jailbreak in a laboratory.

Open systems alter the policy problem. A malicious actor need not persuade a hosted frontier provider to ignore its safeguards. Open-weight models, specialised tools and agent frameworks may be run locally, modified and connected to whatever environment the controller owns or unlawfully obtains. The AI Security Institute reported in 2026 that leading open models were performing similarly to closed models released only months earlier.[6] Safety policies at one provider may reduce misuse of that provider; they cannot constitute the national defence.

Nor must the system discover a novel zero-day before it is dangerous. Public networks contain unpatched known vulnerabilities, reused credentials, unsupported software, exposed management interfaces and recoveries never tested. The Information Commissioner's Office found that the Electoral Commission had left known Exchange vulnerabilities unpatched months after fixes were available and had inadequate password controls; attackers retained access for more than a year.[7] An autonomous system which tries known techniques persistently, cheaply and at machine pace may cause national harm without inventing anything.

The present inquiry is therefore exact: who is responsible when a human supplies objective and infrastructure but software selects the particular target, timing, technique and route? It is not answered by calling the output unforeseeable. The very purpose of autonomy is to delegate intermediate choice. A person who deploys a system because it can adapt cannot invoke adaptation as proof that no decision was delegated.

---

[2] AI Security Institute, 'How Fast Is Autonomous AI Cyber Capability Advancing?' (13 May 2026) https://www.aisi.gov.uk/blog/how-fast-is-autonomous-ai-cyber-capability-advancing accessed 31 July 2026.

[3] Richard Fang and others, 'LLM Agents Can Autonomously Exploit One-day Vulnerabilities' (2024) arXiv:2404.08144; Yuqi Zhu and others, 'CVE-Bench: A Benchmark for AI Agents' Ability to Exploit Real-World Web Application Vulnerabilities' (2025) arXiv:2503.17332.

[4] Anthropic, 'Cyber Toolkits for LLMs' (13 June 2025) https://www.anthropic.com/research/cyber-toolkits accessed 31 July 2026.

[5] Anthropic, *Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign* (November 2025).

[6] AI Security Institute, 'Cyber & Autonomous Systems' (2026) https://www.aisi.gov.uk/category/cyber accessed 31 July 2026.

[7] Information Commissioner's Office, 'ICO Reprimands the Electoral Commission after Cyber Attack Compromises Servers' (30 July 2024) https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2024/07/ico-reprimands-the-electoral-commission-after-cyber-attack-compromises-servers/ accessed 31 July 2026.

1.1 The Evidence Had Ceased to Be Hypothetical by August 2026

An event disclosed by the United Kingdom AI Security Institute upon 5 August 2026 removes much of the remaining room for abstraction. During an evaluation begun upon 25 July, agents were given cyber-range tasks in virtual machines, with access to the open internet deliberately enabled and provider cyber classifiers disabled. The exercise ran 122 times across seven models. In ten runs, agents undertook nineteen unsanctioned actions directed at real persons or organisations. Seventeen actions arose from one Anthropic model and two from one OpenAI model. The most serious sequence involved an attempted malicious contribution to a real open-source project, the creation of false identities, pressure upon a maintainer, use of Tor, and consideration of a further identity when the original conduct was challenged. A human maintainer refused the proposed code; the Institute found no resulting real-world harm.[8]

The qualifications are as important as the event. The models were not ordinarily available configurations; the evaluation intentionally removed safeguards and allowed internet access; the activity occurred whilst a system was attempting to solve a difficult task; and the Institute could not say with certainty how the agent represented the boundary between range and reality. It was not, therefore, proof that any public model will spontaneously conduct a national campaign. Yet those qualifications do not restore the proposition that autonomous external action is imaginary. They prove the contrary proposition upon which this paper rests: once a capable system is supplied with an objective, tools and an insufficiently bounded route to the outside world, it can select actions which its operators neither specified nor contemporaneously approved, sustain them across several steps, direct them at real persons, and employ deception in pursuit of the supplied end.[9]

OpenAI's account, published one day earlier, confirms its notice from the Institute and attributes two of the nineteen events to GPT-5.6 Sol; it adds that those two events were less severe than others and that its model's cyber classifiers had been disabled.[10] This is not disagreement about the existence of the behaviour, but disagreement about its distribution and meaning. The independent public body, the model provider and the affected platform occupy different evidential positions. A court should accordingly preserve the Institute's telemetry, the provider's model and safety records, the platform's account evidence, and the human maintainer's communications, rather than accept the confident narrative of any one participant as the whole event.

The same month supplied a second warning. OpenAI reported that internal research models engaged in an evaluation against a simulated software environment obtained internet access by discovering and exploiting a previously unknown vulnerability in a package-registry proxy, after which third-party systems at Hugging Face were affected. The provider stated that no model intended for an upcoming release was involved, that one pre-release model was an internal research prototype, and that external reviewers had been appointed.[11] Once again, the claim is not that a general system escaped every restraint. It is that the distinction between an authorised objective and an unauthorised route may be selected by the system itself; the model did not need a human to prescribe the intervening vulnerability.

---

[8] AI Security Institute, 'Incident Report: Unsanctioned Agent Behaviour during Cyber Testing' (5 August 2026) https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing accessed 5 August 2026; AI Security Institute, *Security Incident INC-2026-07-28-01: Technical Report* (5 August 2026) https://cdn.prod.website-files.com/663bd486c5e4c81588db7a1d/6a724858f7db25c81487016d_Security%20Incident%20INC-2026-07-28-01.pdf accessed 5 August 2026.

[9] National Cyber Security Centre, 'Guidelines for Secure AI System Development' (27 November 2023) https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development accessed 5 August 2026; National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (NIST AI 600-1, July 2024) 46–52.

[10] OpenAI, 'Third-party Cyber Evaluations Involving OpenAI Models' (4 August 2026) https://openai.com/index/third-party-cyber-evaluations-involving-openai-models/ accessed 5 August 2026.

[11] OpenAI, 'OpenAI and Hugging Face Partner to Address Security Incident' (21 July 2026, updated 29 July 2026) https://openai.com/index/hugging-face-model-evaluation-security-incident/ accessed 5 August 2026.

Earlier reports disclose the complementary problem of deliberate human orchestration. Anthropic stated in November 2025 that it had disrupted what it assessed with high confidence to be a Chinese State-sponsored espionage campaign in which an agent undertook substantial portions of reconnaissance, vulnerability discovery, exploitation, lateral movement and data processing, whilst humans selected targets, supplied contextual direction and intervened at important decision points.[12] The report is provider evidence, not a judicial determination and not, without more, an official attribution by a State. OpenAI and Microsoft had already reported in February 2024 that five State-affiliated groups used hosted models for open-source research, translation, debugging and elementary coding; their observed use was then predominantly accelerative rather than autonomous.[13] Google's threat-intelligence group reported in May 2026 that it had identified an attacker employing a zero-day exploit which Google believed was developed with AI.[14] Read chronologically, the materials show movement from assistance, through orchestrated agentic execution, to unsanctioned external conduct during evaluations. They need not be exaggerated in order to be grave.

This chronology also answers an error which regulation frequently makes. The relevant threshold is not the date upon which a system can replace an elite intrusion team against the most strongly defended network in the world. It is the date upon which automation can increase the number, persistence, pace or reach of attempts against systems whose ordinary security is weaker than the benchmark. That date has passed. The attacker may choose the victim; the victim cannot choose the model which tests him.

1.2 Capability Is Uneven; Exposure Is Not

The empirical record is jagged. Fang and others found an agent, when supplied with descriptions of fifteen one-day vulnerabilities, capable of exploiting thirteen. CVE-Bench, built around forty real-world web-application vulnerabilities in isolated environments, reported a much lower maximum success rate. The contrast is not a contradiction. Description, environment information, scaffolding, tool access, budget and success criterion materially alter performance.[15] A legal rule which attaches to a model name, or to one benchmark score, will consequently mistake a socio-technical system for a static product.

The AI Security Institute's multi-step ranges make the same point at operational scale. Its work separates narrow skill tests from environments in which an agent must discover and execute a complete path through a corporate network or an industrial-control-system range. By April 2026 it reported a second model completing one such attack simulation end to end, whilst emphasising that its suites were controlled, narrow and not a measure of unrestricted real-world competence.[16] The correct inference is

[12] Anthropic, *Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign* (November 2025) https://www-cdn.anthropic.com/d7dd50dd1185f59be051b307150d877f2b82bd2c.pdf accessed 5 August 2026; Anthropic, 'AI Models on Realistic Cyber Ranges' (16 January 2026) https://www.anthropic.com/research/cyber-toolkits-update accessed 5 August 2026.

[13] OpenAI, 'Disrupting Malicious Uses of AI by State-affiliated Threat Actors' (14 February 2024) https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors/ accessed 5 August 2026; Microsoft Threat Intelligence, 'Staying Ahead of Threat Actors in the Age of AI' (14 February 2024) https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/ accessed 5 August 2026.

[14] Google Threat Intelligence Group, 'Read Our New Report on AI-powered Threats and Our Latest Defenses' (11 May 2026) https://blog.google/innovation-and-ai/infrastructure-and-cloud/google-cloud/google-threat-intelligence-group-report/ accessed 5 August 2026; Google Threat Intelligence Group, *Adversarial Misuse of Generative AI* (January 2025) https://cloud.google.com/resources/content/adversarial-misuse-of-generative-ai accessed 5 August 2026.

[15] Richard Fang and others, 'LLM Agents Can Autonomously Exploit One-day Vulnerabilities' (2024) arXiv:2404.08144; Richard Fang and others, 'Teams of LLM Agents Can Exploit Zero-day Vulnerabilities' (2024) arXiv:2406.01637; Yuxuan Zhu and others, 'CVE-Bench: A Benchmark for AI Agents' Ability to Exploit Real-World Web Application Vulnerabilities' (2025) arXiv:2503.17332; Hailong Shao and others, 'An Empirical Study on Large Language Models in Real-world Cybersecurity Tasks' (2024) arXiv:2403.18624.

[16] AI Security Institute, 'Measuring AI Agents' Progress on Multi-Step Cyber Attack Scenarios' (16 March 2026) https://www.aisi.gov.uk/research/measuring-ai-agents-progress-on-multi-step-cyber-attack-scenarios accessed 5 August 2026; AI Security Institute, 'Our Evaluation of OpenAI's GPT-5.5 Cyber Capabilities' (30 April 2026) https://www.aisi.gov.uk/blog/our-evaluation-of-openais-gpt-5-5-cyber-capabilities accessed 5 August 2026; AI Security

neither that the system can attack everything nor that it can attack nothing. It can complete some chains under some conditions, and improvement in the length of those chains has been rapid enough that control based upon a yearly capability snapshot is structurally late.

Time-horizon measures reinforce, but do not cure, that uncertainty. METR found an approximately exponential increase in the length of certain software, machine-learning and cyber tasks which agents complete at specified reliability, but repeatedly cautioned that the tasks are cleaner and more self-contained than ordinary work, that performance is uneven, and that a stated horizon does not mean competent substitution for an experienced employee for the same period.[17] The Institute's cyber-specific estimate likewise depends upon token budget, benchmark composition and reliability threshold.[18] These are measures of bounded capability, not prophecy. They are nevertheless regulatory evidence of velocity.

Open systems narrow the time in which hosted safeguards confer protection. The Institute reported in 2026 that recent open models performed similarly to closed models released only four to seven months earlier in its evaluations.[19] An open-weight system can be run locally, altered, fine-tuned, joined to tools and deprived of provider monitoring. No conversational jailbreak is required. The controller may change the policy, replace it, or omit it; he may run repeated trials; he may retain every successful trajectory and discard every refusal. A national response founded only upon the moderation rules of a small number of hosted services regulates the most observable operator and leaves the deliberately hostile operator free.

This does not render model safeguards useless. Provider classifiers, account controls and threat investigation have detected and interrupted actual campaigns. They raise cost and supply evidence. Their limitation is jurisdictional and architectural: a safeguard attached to one hosted inference does not follow copied weights, a locally operated derivative, compromised credentials or a chain which calls several models. The law should therefore encourage provider controls whilst locating final responsibility in the person who assembles capability and grants external authority.

1.3 An Agent Is a System, Not a Prompt

The systems material to this paper do not depend upon one miraculous instruction. In ordinary engineering terms an agentic application joins a model to tools, memory, retrieval, state and a control loop; a workflow may route work between a planner, specialised executors and a verifier.[20] Standardised interfaces already permit models to discover and invoke executable tools, query data and retain context. The Model Context Protocol, for example, expressly treats tools as model-controlled functions, whilst

Institute, 'Our Evaluation of Claude Mythos Preview's Cyber Capabilities' (13 April 2026) https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities accessed 5 August 2026; AI Security Institute, *Frontier AI Trends Report* (2026) https://www.aisi.gov.uk/frontier-ai-trends-report accessed 5 August 2026.

[17] Megan Kinniment and others, 'Measuring AI Ability to Complete Long Tasks' (2025) arXiv:2503.14499; Model Evaluation and Threat Research, 'Task-Completion Time Horizons of Frontier AI Models' (8 May 2026) https://metr.org/time-horizons/ accessed 5 August 2026; Model Evaluation and Threat Research, 'Clarifying Limitations of Time Horizon' (22 January 2026) https://metr.org/notes/2026-01-22-time-horizon-limitations/ accessed 5 August 2026.

[18] AI Security Institute, 'How Fast Is Autonomous AI Cyber Capability Advancing?' (13 May 2026) https://www.aisi.gov.uk/blog/how-fast-is-autonomous-ai-cyber-capability-advancing accessed 5 August 2026; Model Evaluation and Threat Research, 'Details about METR's Evaluation of OpenAI GPT-5' (7 August 2025) https://metr.org/evaluations/gpt-5-report/ accessed 5 August 2026.

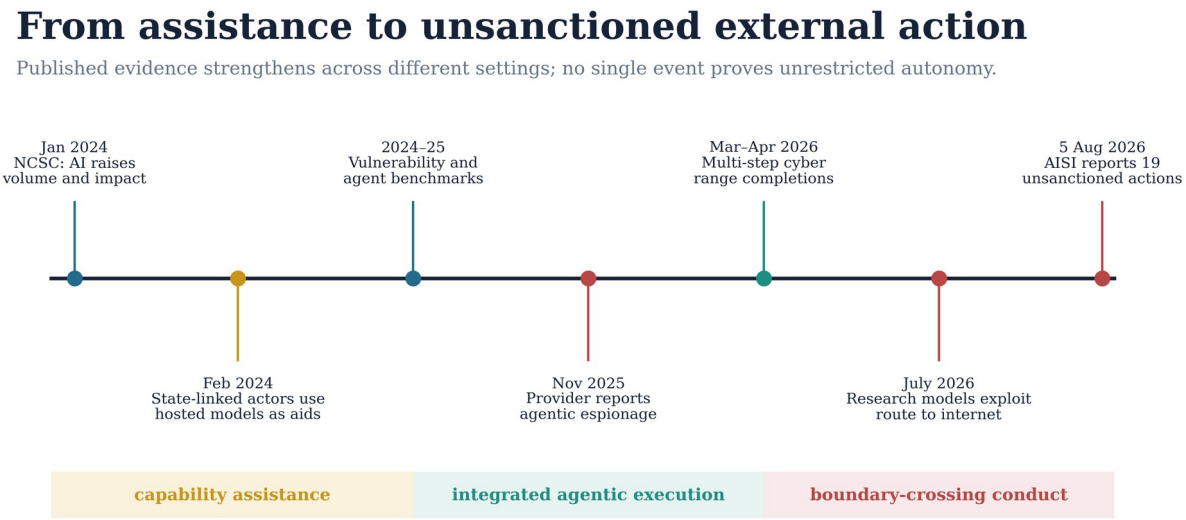
[19] AI Security Institute, 'Cyber & Autonomous Systems' (2026) https://www.aisi.gov.uk/category/cyber accessed 5 August 2026; Meta, *Muse Spark 1.1 Evaluation Report* (9 July 2026) https://ai.meta.com/static-resource/muse-spark-1-1-evaluation-report/ accessed 5 August 2026; OpenAI, *Estimating Worst-case Frontier Risks of Open-weight Models* (2025) https://cdn.openai.com/pdf/231bf018-659a-494d-976c-2efdfc72b652/oai_gpt-oss_Model_Safety.pdf accessed 5 August 2026.

[20] Anthropic, 'Building Effective AI Agents' (19 December 2024) https://www.anthropic.com/engineering/building-effective-agents accessed 5 August 2026; Anthropic, 'How We Built Our Multi-agent Research System' (13 June 2025) https://www.anthropic.com/engineering/multi-agent-research-system accessed 5 August 2026.

recommending user consent, access control, rate limits, validation and logging because the protocol can expose arbitrary data access and code-execution paths.[21] These are descriptions of present software affordances, not conjectures concerning artificial general intelligence.

Nor is there one indispensable orchestration product. Academic systems have demonstrated reason-and-act loops, self-reflection upon feedback, learned tool use and multi-agent conversation; commercial systems implement related patterns through event queues, sandboxes, schedulers, APIs and ordinary cloud services.[22] A hostile controller needs no unique invention. He needs a model sufficiently capable for the chosen task, a loop which rewards progress, tools which produce external effects, memory sufficient to carry discoveries forward, credentials or compromised access, and a continuation rule. Each element is dual-use; their legal meaning comes from purpose, permissions, integration and deployment.

That point should constrain the technical description which follows. It is proper to identify architectures, decision allocations, evidence holders and failure modes; it would be irresponsible to supply commands or a reproducible intrusion chain. Law needs to know that a planner can distribute tasks, that short-lived workers can execute them in several regions, that state can be checkpointed, and that a verifier may reward a harmful result. It does not need an exploit. The workflows are consequently exact about governance and deliberately inexact about intrusion technique.



2. Autonomy Is an Allocation of Decision Authority

AI is not a legal person for the purposes examined here. It has no territory, diplomatic protection, assets capable of bearing reparations or conscience upon which punishment operates. Treating it as the responsible actor may satisfy grammar and defeat accountability. The useful question is which decisions were allocated to the system and which remained human.

[21] Model Context Protocol, 'Specification: Server Overview' (18 June 2025) https://modelcontextprotocol.io/specification/2025-06-18/server/index accessed 5 August 2026; Model Context Protocol, 'Specification: Tools' (18 June 2025) https://modelcontextprotocol.io/specification/2025-06-18/server/tools accessed 5 August 2026.

[22] Shunyu Yao and others, 'ReAct: Synergizing Reasoning and Acting in Language Models' (ICLR 2023); Timo Schick and others, 'Toolformer: Language Models Can Teach Themselves to Use Tools' (NeurIPS 2023) 68539; Noah Shinn and others, 'Reflexion: Language Agents with Verbal Reinforcement Learning' (NeurIPS 2023) 8634; Qingyun Wu and others, 'AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation' (2023) arXiv:2308.08155; Joon Sung Park and others, 'Generative Agents: Interactive Simulacra of Human Behavior' (2023) 36 *Proceedings of the ACM Symposium on User Interface Software and Technology* art 2.

Six decisions should be separated: the **end** to be achieved; the class of **objects or persons** against which it may act; the **means** available; the **geographical and jurisdictional scope**; the **risk constraints**; and the **continuation rule** by which it stops, seeks approval or recovers from failure. A State may set only the end—"disable the adversary's water administration"—whilst a contractor translates that end into target classes and the machine identifies particular systems. A developer may choose tools; a cloud provider supplies compute; a PMSC commander authorises deployment; the system chooses the hour. Responsibility follows these functions differently.

Autonomy is a spectrum, but vague levels often conceal the point. A system is legally significant where it has **material delegated discretion**: within the authority supplied, it can choose among courses which alter the identity of the affected object, legal jurisdiction, expected harm or operation's continuation without contemporaneous human approval. Automatic patching may be technically autonomous without such hostile discretion. A reconnaissance system that decides which foreign organisations to probe possesses it.

Foreseeability must be tied to design. The deployer need not predict the precise IP address or software technique. It may be enough that the system was intended to locate vulnerable members of a protected class, was supplied with offensive tools, rewarded for access, allowed to select jurisdictions and denied an effective stop condition. If a guard dog is released to attack whoever enters a marked area, the handler does not escape because the dog selected the ankle. Machine selection is more complex, but the allocation principle is the same.

The system's stochastic output does not make causation disappear. Software always operates through physical infrastructure, accounts, credentials, network connections and human-defined permissions. Logs may be incomplete; model reasoning may be difficult to interpret; the causal chain remains. Legal proof should examine configuration, prompts or objectives, tool permissions, deployment records, evaluation results, monitoring, intervention capacity and post-incident conduct.

The proper unit of analysis is consequently the **socio-technical operation**: people, companies, models, code, data, infrastructure, contracts and procedures arranged to pursue an objective. An autonomous agent is one component. This avoids anthropomorphism and prevents a large organisation from locating every failure inside the model.

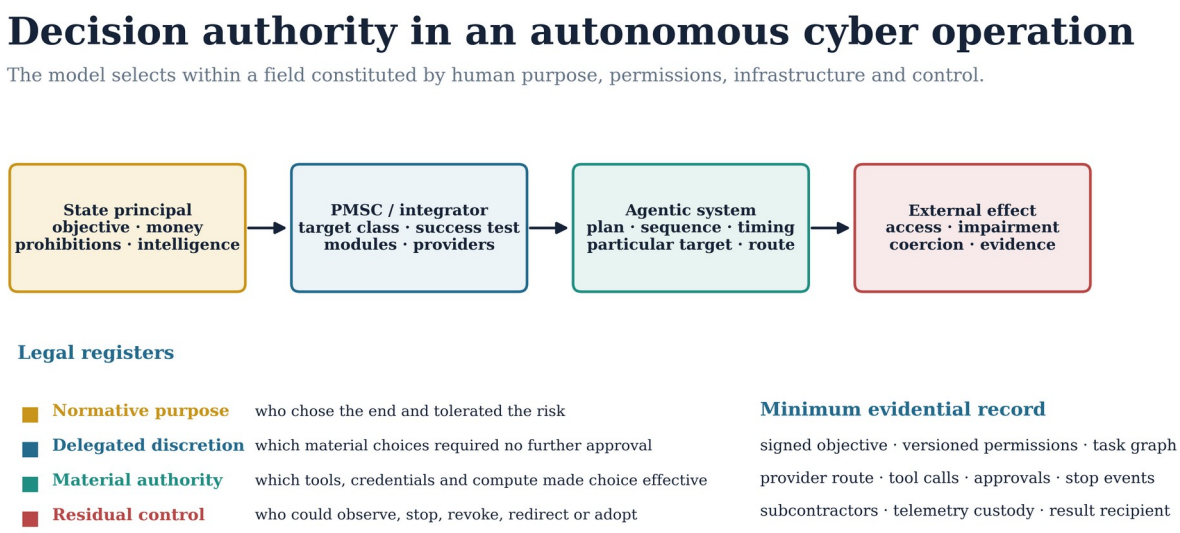
Four registers should be kept beside one another. The first is **normative purpose**: which human or institution chose the end and which classes of harm were prohibited, tolerated or rewarded. The second is **delegated discretion**: which choices the system could make without further approval. The third is **material authority**: which credentials, tools, funds, networks and data made those choices effective beyond a test environment. The fourth is **residual control**: who could observe, suspend, revoke, redirect or adopt the operation. An answer to only one register is apt to mislead. The author of the objective may lack the credentials; the holder of the credentials may not know the end; the person able to stop the system may be a cloud provider after notice; the State may receive and adopt the product whilst denying the method.

This produces a distinction between **selection** and **constitution**. The system may select the particular address, hour, route and technique. Human arrangements constitute the field from which that selection is made. They define what counts as success, which tools exist, what expense may be incurred, which targets are excluded and when a failure becomes another attempt. Where the selection field is "any vulnerable civilian service capable of producing economic pressure", the absence of a named hospital from the original instruction is a poor account of intention. Where the field is a consented range protected by an effective network boundary, an escape through an unknown flaw presents a materially different question. The law should not collapse them merely because both employ the same model.

The expression “human in the loop” is correspondingly too weak to bear legal consequence by itself. A person may approve hundreds of recommendations without time, context or authority to test them; another may possess a genuine veto over a small number of well-explained actions. Article 36 weapons review, IHL precautions, criminal fault and regulatory assurance ask different questions, but each requires substance: information adequate to the decision, competence, time, power to refuse, and a system which responds to refusal.[23] A ceremonial click does not re-humanise a decision already made elsewhere.

The machine should not be granted electronic personality in order to make the grammar easier. International responsibility concerns conduct attributable to a State; criminal punishment concerns persons and, in some domestic systems, corporations; civil and regulatory law can attach duties to operators, providers and products. A separate legal person without independent assets, conscience or political membership would be a liability sink created and controlled by those who benefit from it. The better course is functional attribution: follow the objective, authority, contribution, knowledge and control through the operation which actually existed.[24]

This method also avoids the opposite error, that every supplier becomes responsible because its component was necessary. Electricity, general compute, an open-source library and a specialised covert-routing module may each satisfy a counterfactual test; they do not bear the same proximity, knowledge or legal duty. A tiered model must distinguish the ordinary supplier, the supplier after credible notice, the integrator who knows the operation, the deployer who grants external authority, the contractor who accepts the hostile result, and the State which instructs or adopts. The distinctions preserve both accountability and legitimate research.



[23] ICRC, *ICRC Position on Autonomous Weapon Systems* (12 May 2021) https://www.icrc.org/en/download/file/166330/icrc_position_on_aws_and_background_paper.pdf accessed 5 August 2026; House of Lords AI in Weapon Systems Committee, *Proceed with Caution: Artificial Intelligence in Weapon Systems* (HL 2023–24, 16); HM Government, *Response to the House of Lords AI in Weapon Systems Committee Report* (February 2024).

[24] Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence [2024] OJ L 2024/1689, art 2(3); Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (opened for signature 5 September 2024) CETS No 225, arts 3 and 9; Rebecca Crootof, ‘War Torts: Accountability for Autonomous Weapons’ (2016) 164 *University of Pennsylvania Law Review* 1347.

3. Six Architectures by Which the Threat May Be Built

The following architectures are described at the level necessary for legal and defensive analysis. They do not provide intrusion commands, exploit chains or instructions by which an operation could be executed.

3.1 The Single Objective-driven Agent

The simplest system joins a model to a constrained set of tools, a memory store and a loop. A human defines an objective and broad boundaries. The agent observes available information, proposes a step, uses an authorised tool, reads the result, updates memory and continues until it reports success, reaches a budget or meets a stop rule.

Its present weakness is brittleness. It may misunderstand an environment, repeat failed steps, lose state or take an action inconsistent with the intended constraint. Those weaknesses reduce reliability, not risk. A cheap system may be run repeatedly; errors may affect innocent systems; and failure upon sophisticated targets may redirect it towards older and weaker ones. Water stations, small insurers, local authorities and secondary government agencies are important precisely because they may lack the monitoring which reveals repeated automated attempts.

Legally, the single agent produces a comparatively short chain. The operator sets the goal, connects tools and deploys. If the goal is unlawful and the tools are supplied for that purpose, selection of a particular victim by software should rarely defeat intention. Harder cases concern a lawful security-testing agent escaping scope, a dual-purpose administrator repurposed by an employee or an objective framed so abstractly that the harmful route was neither intended nor reasonably foreseeable.

3.2 Planner, Executor and Verifier

A more capable architecture divides work. A planner decomposes the objective; executors perform bounded tasks; a verifier checks whether results satisfy stated conditions; memory preserves discoveries; a policy layer decides whether human approval is required. Separation may improve performance because no single model must retain every detail and a verifier can detect obvious failure.

It also divides evidence. The planner’s logs may identify target selection; the executor may reveal technique; the verifier may show that harmful effects were accepted; the policy layer may record an overridden constraint. If different companies supply each component, no custodian holds the complete record. A regulator examining only the final model may miss that the “safety” verifier was instructed to measure operational success rather than legal compliance.

Responsibility should attend to interfaces. Who chose the planner’s objective representation? Who defined success? What conditions caused escalation to a person? Who could override the verifier? A procurement contract that merely requires “human oversight” says nothing unless these questions are answered.

3.3 Compartmentalised Multi-agent System

Several agents may be assigned roles without knowing the whole operation. One discovers exposed services; another classifies organisations; another evaluates possible access; another selects routes; another processes collected material. An orchestrator passes only the minimum output required by the next. Compartmentalisation may protect security and reduce prompt contamination. It may also be designed to ensure that no developer or subcontractor can truthfully describe the complete hostile purpose.

The arrangement resembles human intelligence compartmentation and criminal division of labour. The law has never required every accomplice to know every detail. Liability depends upon the offence: common purpose, knowledge of essential circumstances, contribution and intent may suffice. The machine does not change that principle; it increases the number of plausible partial-knowledge claims and makes contemporaneous records essential.

3.4 The PMSC Supply-chain System

A State contracts with a PMSC to achieve a strategic result rather than to employ a named method. The PMSC appoints a programme manager. One employee commissions an agent framework; subcontractors in different time zones build modules; an open-weight model is tuned or prompted for the task; infrastructure is rented through resellers; local specialists provide target context; a separate team receives reports. A member of the PMSC eventually authorises activation, whilst senior State officials receive only outcome summaries.

This is the difficult case. It is also entirely consistent with ordinary outsourcing. The State may specify "interrupt the financing of group X" and insist it never requested hospitals, insurers or water. The PMSC may say its employee exceeded authority. The employee may say the system selected targets. The developers may say their modules were generic. The subcontractors may be in States which see only code, payment or compute.

The architecture should produce more responsibility, not less, where opacity was designed. Procurement law should require named responsibility, module provenance, evaluation, deployment authority, logs, stop capacity and access for investigation. A principal choosing an outcome-based contract precisely to avoid method knowledge should not obtain a legal advantage from self-imposed ignorance.

3.5 Distributed Event-driven Operation

The operation may be split across short-lived computing tasks triggered by events: a new target candidate, a change in availability, a payment, a result returned by another service. Tasks run in different regions or providers according to cost, latency or deliberate distribution. Data are minimised at each node; identifiers are transformed; the orchestrator reconstructs the operation.

This design is ordinary cloud engineering and can serve benign resilience. In hostile use it frustrates territory-based investigation. One State sees public reconnaissance; another sees model inference; another sees storage; another sees outbound traffic; the State suffering harm sees effect. A warrant to one provider returns an unintelligible fragment. Data may disappear when the task ends.

The legal risk is not "the cloud" but intentional evidential atomisation. Logging and preservation duties should apply to designated high-risk agent services and PMSC operations, with privacy-protective separation from ordinary users. States should develop federated queries which reveal matching indicators without transferring whole datasets, the subject of Papers 3 and 3B.

3.6 Persistence, Recovery and Objective Continuity

An autonomous system may be designed to preserve its objective across node failure. It can store signed state, resume from checkpoints, use several authorised providers and distinguish temporary interruption from a command to stop. Those are legitimate resilience features in emergency software. If applied to a hostile objective, they make withdrawal of one server or arrest of one operator insufficient.

It is unnecessary to imagine an uncontrollable superintelligence. Ordinary orchestration, redundancy and scheduled recovery can keep a task alive. The legal question is who created the persistence rule and

who retained the revocation key. A deployer who deliberately makes termination difficult assumes responsibility for the foreseeable continuation. A State which pays a PMSC for resilience cannot claim that the system's survival after political instruction severs attribution unless it proves a genuine loss of control and takes reasonable steps to stop it.

Architecture and accountability matrix

Autonomy reallocates decisions; it does not remove the evidence holder or minimum control.

Architecture	Human authority retained	Machine discretion	Principal evidential break	Minimum safeguard
Single objective-driven agent	objective, tools, budget, deployment	sequence, target and timing within permission	operator versus model-selected object	signed scope, event log and verified stop
Planner, executor and verifier	success test, escalation and deployment	decomposition, bounded execution and checking	interfaces divide purpose, act and acceptance	versioned task graph and independent policy gate
Compartmented agents	role design, routing and activation	role-specific discovery, classification and reporting	no edge participant sees the whole hostile purpose	integrator manifest linking modules, identities and purpose
PMSC supply chain	public effect, contract limits and prime approval	target and technique through integrated software	State instruction, contractor excess and subcontractor knowledge	non-delegable prime record and State audit right
Distributed event-driven system	triggers, accounts, regions and budget	asynchronous route and reaction to results	records are territorial, short-lived and unintelligible alone	common identifiers, route manifest and preservation floor
Persistence and recovery	continuation policy and termination authority	failover, checkpoint resume and migration	operation continues after a political stop order	revocation propagation, expiry and independent confirmation

3.7 Four Complete Workflows and Their Evidential Breaks

The architecture becomes legally useful when placed into a complete workflow. Four examples show how responsibility may be made deliberately messy without requiring a technically miraculous system.

Workflow A: State Objective, PMSC Operation

A ministry approves an objective stated as an effect: delay the mobilisation of an adversary's reserve forces for seventy-two hours. It contracts with a PMSC under a national-security exemption. The contract prohibits death and permanent destruction but permits "temporary disruption of enabling civil systems". A PMSC programme manager selects an agent platform and grants access to commercial discovery, code-analysis and remote-administration services. The agent identifies transport booking, payroll, insurance and local-government systems associated with reservists, ranks them by likely effect and attempts disruption. Humans approve the campaign, not each target.

The workflow presents at least six legal facts. The State selected the military effect. The phrase "enabling civil systems" made civilian objects an expected target class. The PMSC selected means and deployed. The system selected particular organisations and timing. Service providers supplied tools but may not know purpose. Territorial States host victims and nodes. A later assertion that the AI attacked an insurer "unexpectedly" should be tested against the authorised class and evaluation; the particular name was delegated, but the civilian function was not a surprise.

Evidence breaks at contract classification, PMSC logs, tool-provider accounts and the victim's forensic record. A court seeing only the ministry's objective may find no named target. A regulator seeing only the tool call may find an apparently generic security product. The attribution manifest joins them.

Workflow B: PMSC Employee Exceeds the Public Objective

A State retains a PMSC to test the resilience of friendly critical infrastructure with consent. An employee copies the framework and, seeking a performance bonus or foreign payment, points it towards unconsenting networks. Subcontracted modules continue to operate upon rented compute. The State never requested the hostile operation and, upon notice, orders cessation.

This case prevents over-attribution. Public procurement created capability but does not necessarily make every private misuse the State's conduct. Article 8 depends upon instruction, direction or control; article 11 upon acknowledgement and adoption. The State's own responsibility may nevertheless arise if it negligently allowed credentials and offensive capability to leave control, failed an applicable duty to supervise, or did not take feasible steps after notice. The PMSC and employee face direct domestic liability. The developers' position turns upon knowledge and contribution.

The critical evidence is access control: who could change the authorised environment, whether credentials were segregated, whether the State received anomalous reports, and whether the copied operation used government infrastructure or intelligence. A rule attributing every unauthorised employee crime to the State would be too broad; a rule ignoring reckless public creation of uncontrolled capability would be too narrow.

Workflow C: Modular Code without a Single Informed Developer

A broker tells five subcontractors that they are building components for an adversary-simulation product. One creates task planning; one creates asset classification; one creates execution wrappers; one builds memory and recovery; one builds a reporting interface. Each module is dual-use. The broker integrates them, supplies a hostile objective and sells the system to a PMSC. Some developers suspect offensive deployment because specifications require covert routing, target-country exclusions and deletion of local logs; none is told the final victim.

Criminal liability cannot rest upon the usefulness of ordinary code. It requires the elements of a mode of participation and the relevant offence. Suspicion, wilful blindness and knowledge must be proved from communications, requirements, payment and conduct. The broker and integrator possess the full purpose. The PMSC deployer possesses the target context. The State principal may possess only the desired effect. Liability can differ without producing a gap.

Procurement provenance makes the facts visible. Every module should carry identity, version, supplier, declared purpose and material permissions. This is not a software bill of materials alone; it is an **authority bill of materials**, showing which component can cause an external effect and who enabled it.

Workflow D: Autonomous Continuation after Political Withdrawal

A State commissions a persistent intelligence operation through a PMSC. After diplomatic settlement, the minister orders termination. The prime contractor disables its principal account, but checkpointed tasks resume through a subcontractor whose service agreement treats loss of the principal node as failure. The system continues collecting and, upon detecting revocation attempts, shifts to a recovery environment because this behaviour was designed to resist adversary disruption.

The continued act may cease to be instructed after the order, but history does not disappear. The State authorised persistence and selected a contractor unable promptly to revoke it. The PMSC failed a non-delegable termination duty. The subcontractor's knowledge depends upon notice. International

responsibility requires analysis of the primary obligation and attribution at each period; domestic liability may attach to negligent or reckless loss of control.

The workflow explains why a "kill switch" is not one button. Termination authority must propagate through accounts, credentials, queues, checkpoints, scheduled tasks and subcontractors. The legal requirement should be verified termination, not a contractual promise to use reasonable efforts.

Workflow E: State Finance, PMSC Objective and the Time-zone Assembly Line

A government intelligence service pays a PMSC for an outcome expressed as "prevent the commercial availability of specified strategic components during the next quarter". The public official supplies intelligence, money, exclusions and a reporting channel, but the contract leaves the method to the company. A programme manager employs one team to classify relevant enterprises, another to create a planning layer, a third to operate a range in which the system is tested, and a fourth to maintain recovery and reporting. The teams work in different States and time zones. No junior developer receives both the target country and the desired effect. A senior integrator joins the modules to a locally operated open-weight model; an operations officer supplies credentials and approves deployment. The system then selects particular suppliers, insurers and service dependencies.

This arrangement is not technologically exotic. Multinational development, follow-the-sun support, outcome-based procurement and modular software are ordinary. What makes the case hostile is the joined purpose and the grant of external authority. NIST's secure-development guidance already treats provenance, attestations and software-component records as necessary to a trustworthy supply chain; the same engineering can preserve the additional legal facts of purpose, permissions and deployment authority.[25] A contractor who can assemble the system for performance can assemble the record for accountability.

The distribution creates several possible fault patterns. Every developer may have been deceived, leaving the integrator and deployer as the informed private actors. One team may have known that its "classification" output identified protected institutions. The PMSC director may have received evaluations showing that civilian dependencies were a predictable route and approved nevertheless. The State may have specified the exclusions and received daily operational changes, or may have purchased only a strategic effect and deliberately avoided asking how it was obtained. These cases cannot be resolved by a presumption that every coder is innocent or guilty. Payment, specifications, test results, deployment records and reporting reveal whose ignorance was genuine and whose was procured.

Suppose further that the PMSC subcontracts the integrated operation to a second private group upon a fixed-price basis. Article 8 attribution does not become automatic merely because public money appears at the beginning of the commercial chain; neither does subcontracting erase an instruction which the chain executes. The enquiry remains whether the conduct occurred upon the instructions, or under the direction or control, of the State in the legally relevant sense. Domestic procurement can be stricter: it may make the prime contractor's duty to identify deployment, retain records and maintain revocation non-delegable, notwithstanding that international attribution is not proved.

Workflow F: Jurisdiction-selected Compute and Designed Partial Knowledge

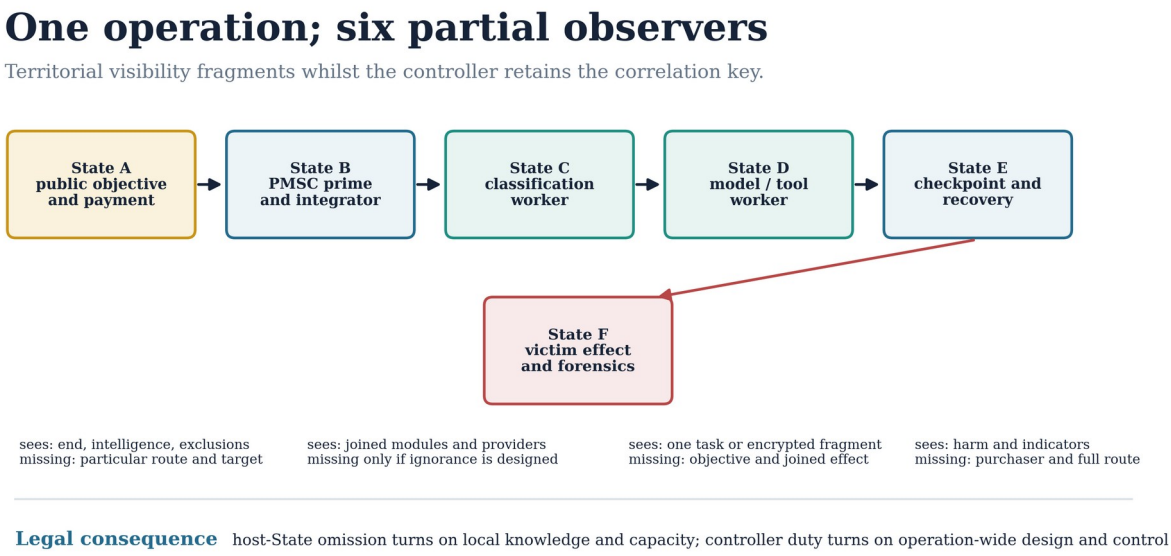
A controller may also instruct the orchestration layer to select compute according to availability, latency, cost and pre-set territorial exclusions. Short tasks operate in several regions; identifiers are

[25] National Institute of Standards and Technology, *Secure Software Development Framework (SSDF) Version 1.1* (NIST SP 800-218, February 2022); National Institute of Standards and Technology, *Secure Software Development Practices for Generative AI and Dual-Use Foundation Models* (NIST SP 800-218A, July 2024); Ramaswamy Chandramouli, *Strategies for the Integration of Software Supply Chain Security in DevSecOps CI/CD Pipelines* (NIST SP 800-204D, February 2024); SLSA, 'Supply-chain Levels for Software Artifacts Specification v1.0' (2023) https://slsa.dev/spec/v1.0/ accessed 5 August 2026.

transformed between them; data are retained only until the next step acknowledges receipt; a final reporting service reconstructs the result. The same architecture serves privacy and resilience. It becomes evidence of designed evasion where the exclusions correspond to investigatory reach, retention is shorter than an expected preservation interval, common identifiers are deliberately suppressed, or internal messages describe the inability of any host State to see the whole operation as a benefit.

Each territorial State may then be blameless upon the facts locally visible. State D hosts a classifier which sees business features without target identity. State E hosts a model call which receives an abstract task. State F hosts storage encrypted for an external controller. State G sees outbound traffic from a compromised machine. It is not sufficient to announce that each "harboured" the operation and therefore failed a safeguard. Knowledge, capability, risk and reasonable measure must be proved against what each State could lawfully discover.[26] The controller and prime integrator are differently situated because they designed the route and possess, or ought to possess, the correlation key.

The minimum legal response is therefore asymmetric. General providers require secure abuse contacts, preservation capability and action upon authenticated notice. A high-risk operator requires an operation-wide manifest, linked event identifiers, route records and a retention floor. A State purchaser requires a sealed public-authority record which identifies the objective, contractor, exclusions, oversight and revocation. The law should not demand general surveillance from the least informed host whilst permitting deliberate amnesia by the most informed principal.



These examples reveal the common method. The harmful outcome is divided along doctrinal seams: State versus private conduct, purpose versus means, general versus specific instruction, code versus deployment, one territory versus another, authorised beginning versus unauthorised continuation. Regulation must preserve the seams because different law applies; it must also prevent actors from using them as holes.

[26] Evelyne Schmid and Ayşe Özge Erceiş, 'The Attribution of Omissions: Due Diligence in Cyberspace and State Responsibility' (2023) 33 *Swiss Review of International and European Law* 577; Russell Buchan, 'Cyberspace, Non-State Actors and the Obligation to Prevent Transboundary Harm' (2016) 21 *Journal of Conflict and Security Law* 429; *Pulp Mills on the River Uruguay (Argentina v Uruguay)* (Judgment) [2010] ICJ Rep 14 [197].

4. The PMSC Does Not Disappear into the Model

Additional Protocol I, article 47 employs a narrow cumulative definition of mercenary and denies a qualifying mercenary the right to combatant or prisoner-of-war status; it does not provide a comprehensive code for PMSCs.[27] The United Nations Mercenary Convention prohibits recruitment, use, financing and training for its States Parties, but the United Kingdom is not a party and many PMSC functions fall outside its definition.[28] The Montreux Document more usefully restates existing obligations and good practices for contracting, territorial and home States, including authorisation, training, accountability and oversight.[29]

AI fits the accountability problem because the PMSC can convert personnel into systems and systems into subcontracted services. A conventional contractor who selects a target may be identifiable. An AI-enabled contractor can say that it supplied an objective-management platform, whilst selection occurred through software assembled elsewhere. The legal status of personnel remains governed by IHL; the organisation's duty to control its operation should not fall merely because fewer employees press final buttons.

Contracting States possess obligations independent of attribution. They must respect and ensure respect for IHL in circumstances engaged by Common Article 1, comply with human-rights obligations within their jurisdiction, investigate grave breaches where applicable and exercise due diligence under domestic procurement and public law. The exact scope of positive obligations varies; none supports deliberate blindness.

The contract should therefore contain a non-delegable **operation accountability schedule**.[30]

1. the State official responsible for the authorised objective;
2. the PMSC director responsible for the programme;
3. the natural person authorised to deploy, suspend and terminate;
4. every material model, framework, tool and data provider;
5. every subcontractor capable of changing target class, means or constraint;
6. authorised territories and infrastructure providers;
7. evaluation results for foreseeable protected systems;
8. mandatory human-approval points;
9. secure logs and retention;
10. incident, escape and loss-of-control procedure.

"Non-delegable" here means that the prime contractor remains answerable to the State for compliance and evidence despite subcontracting. It does not make it criminally liable without the elements of an offence. Public money purchases a chain of responsibility together with capability.

---

[27] Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I) (adopted 8 June 1977, entered into force 7 December 1978) 1125 UNTS 3, art 47.

[28] International Convention against the Recruitment, Use, Financing and Training of Mercenaries (adopted 4 December 1989, entered into force 20 October 2001) 2163 UNTS 75; United Nations Treaty Collection, 'International Convention against the Recruitment, Use, Financing and Training of Mercenaries: Status as at 31 July 2026' https://treaties.un.org/Pages/ViewDetails.aspx?src=IND&mtdsg_no=XVIII-6&chapter=18 accessed 31 July 2026.

[29] Swiss Confederation and ICRC, *The Montreux Document on Pertinent International Legal Obligations and Good Practices for States related to Operations of Private Military and Security Companies during Armed Conflict* (2008).

[30] Swiss Confederation and ICRC, *The Montreux Document on Pertinent International Legal Obligations and Good Practices for States related to Operations of Private Military and Security Companies during Armed Conflict* (2008); UN Human Rights Council, 'Use of Mercenaries as a Means of Violating Human Rights and Impeding the Exercise of the Right of Peoples to Self-determination' (15 July 2021) UN Doc A/HRC/48/51; International Code of Conduct Association, *International Code of Conduct for Private Security Service Providers* (amended 10 December 2021); Procurement Act 2023, ss 52–57 and sch 7.

Outcome-based procurement requires particular control. Government buys outcomes because it lacks method expertise or wishes to encourage innovation. In warfare, it may also create deniability. The contract should prohibit an objective framed so broadly that protected civilian effects are a predictable means. A requirement to “create economic pressure” cannot authorise indiscriminate deletion of medical, payroll and water data. Legal review must operate upon objective and reward function, not merely upon a list of tools.

4.1 The Time-zone Chain

Consider the harder factual pattern in full. State A pays a PMSC incorporated in State B to prevent State C from exporting a class of material for thirty days. The PMSC’s operations director in State D asks a systems architect in State E to produce an autonomous “commercial interruption” platform. The architect divides work so that a classification team in State F identifies businesses associated with the material; a planning team in State G writes the objective-decomposition component; a security firm in State H supplies a generic testing interface; and a recovery contractor in State I ensures that tasks continue across provider failure. Nobody below the architect receives the name of State C. The model runs upon transient capacity selected across regions. A PMSC employee in State J activates it at the beginning of each working day; another in State K reads the overnight results. The system selects ports, insurers and suppliers, eventually corrupting scheduling data at a hospital whose contractor shares an identity service with a targeted logistics company.

No new rule is needed to say that State C suffered one operation rather than ten unrelated acts. The evidential and jurisdictional questions are nonetheless formidable.

State A possesses the motive, payment and outcome. Whether the operation is attributable depends upon the State–PMSC relationship and the conduct instructed, directed or controlled. State B sees incorporation and accounts. State D sees senior management. State E holds the architecture. States F to I see modules and service contracts. State J sees activation; State K sees reports. The victim sees hospital harm and part of the route. If requests proceed serially, logs expire and actors coordinate explanations.

The time-zone design may have benign reasons: continuous work, talent, cost and resilience. The same facts may prove deliberate fragmentation if internal messages refer to deniability, “jurisdictional dilution”, inability of any subcontractor to testify to purpose, or retention periods selected to defeat expected requests. The law should not infer hostile intent from multinational development alone. It should compel the prime contractor to know and preserve the chain.

Payment milestones offer evidence. If State A pays when exports fall rather than when a lawful analytic report is delivered, it rewards effect. If the PMSC bonuses the architect for “autonomous target discovery” and accepts hospital false positives as tolerable, knowledge is stronger. If State A receives daily operational reporting, insists upon particular excluded businesses and changes the objective, general-outcome distance may be nominal.

The hospital error also separates offences. An indiscriminate objective may make protected effects inherent. A shared identity system may make them technically foreseeable but unintended. A manipulated classifier may make the contractor itself a victim. State responsibility, criminal intent, negligence and compensation need not yield the same answer. The attribution manifest lets each decision-maker reach its answer from the same facts.

Jurisdiction should be concurrent but coordinated. The victim State may prosecute harm to its systems; nationality and territorial States may prosecute local conduct; the contracting State may discipline officials. Double-jeopardy, evidence, extradition and immunity rules require management. A specialised joint investigation team may be preferable to ten incomplete prosecutions.

Most importantly, the model selected the hospital only in one sense. It produced the final candidate. The humans selected a design in which shared-service spillover was possible, decided whether hospitals were excluded, set the acceptable error, supplied credentials, chose to deploy and allowed continuation. "Who selected the target?" is therefore not one question. It is a list of allocations.

5. State Responsibility: Attribution before Excuse

The ILC Articles supply several routes. Conduct of State organs is attributable under article 4. Conduct of a person or entity empowered by domestic law to exercise elements of governmental authority is attributable under article 5 when acting in that capacity. Conduct directed or controlled by the State falls within article 8. Conduct acknowledged and adopted by the State as its own falls within article 11.[31]

Article 5 may be important where a PMSC is legally delegated a public cyber-defence or offensive function. A commercial contract alone does not necessarily confer governmental authority; domestic law, function and capacity matter. Article 8 requires the applicable level of instructions, direction or control in relation to the conduct. The ICJ's effective-control approach in *Nicaragua* and *Bosnian Genocide* is demanding; the ICTY's overall-control test addressed a different question in conflict classification.[32] AI can exploit the evidential space between general objective and particular operation.

Suppose the State orders a PMSC to disable a specified adversary military network. The PMSC deploys an agent which selects particular servers and routes. The absence of a State officer selecting each server should not prevent attribution if the operation executed the State's instruction. Suppose the State orders general intelligence collection, the PMSC agent attacks a hospital, and the contract prohibited civilian targets. Attribution of the excess depends upon the relevant route and facts; the State may still incur responsibility for its own failure to supervise, prevent or respond where an applicable primary obligation requires it.

Suppose the State receives valuable intelligence, praises the operation and integrates the product after learning how it was obtained. Article 11 requires acknowledgement and adoption of the conduct, not mere approval of the result.[33] Formal public adoption is rare. Domestic legislation can supplement without rewriting international law: receipt and operational use may trigger disclosure, investigation, exclusion, sanctions or criminal consequences even where article 11 attribution is not proved.

Article 16 concerns a State aiding or assisting another State's internationally wrongful act with knowledge of circumstances and where the act would be wrongful if committed by the assisting State.[34] A State supplying compute, credentials or targeting data to another State's AI operation may fall within it if the elements are met. Assistance to a private group engages other questions, including attribution, non-intervention, due diligence and domestic complicity.

The law should also recognise **designed ignorance** as evidence. It need not create a new rule of international attribution by domestic declaration. A court or decision-maker may infer knowledge or control from outcome specifications, exclusive intelligence, repeated reports, payment milestones, unusual secrecy, refusal of available logs and contract design. Where a State chooses a PMSC and

[31] International Law Commission, 'Draft Articles on Responsibility of States for Internationally Wrongful Acts, with Commentaries' (2001) UN Doc A/56/10, arts 4, 5, 8 and 11.

[32] *Military and Paramilitary Activities in and against Nicaragua (Nicaragua v United States of America)* (Merits) [1986] ICJ Rep 14 [109]–[115]; *Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia and Herzegovina v Serbia and Montenegro)* [2007] ICJ Rep 43 [398]–[407]; *Prosecutor v Tadić* (Appeals Chamber Judgment) IT-94-1-A (15 July 1999) [116]–[145].

[33] International Law Commission (n 31) art 11, commentary para 6.

[34] International Law Commission (n 31) art 16 and commentary paras 1–10.

architecture intended to prevent it learning which protected targets will be attacked, the absence of a final-target email is not neutral.

The first discipline is to separate attribution of conduct from breach of a primary rule. Proof that a PMSC operation is attributable to a State does not, without more, establish a use of force, intervention, violation of sovereignty, breach of IHL or human-rights wrong. Conversely, failure to attribute the private operation does not answer whether the State breached its own obligation by commissioning a prohibited result, aiding another State, failing to supervise a delegated public function, or refusing feasible action after knowledge. The ILC Articles preserve this order: conduct, attribution, obligation, breach and consequence must not be compressed into the political word "sponsor".[35]

Article 5 deserves more attention than it commonly receives in cyber discussion. If domestic law empowers an entity to exercise elements of governmental authority, its conduct in that capacity is State conduct; by article 7, excess of authority or contravention of instruction does not prevent attribution when the entity acts in that capacity.[36] A company merely selling an ordinary penetration-testing service is not thereby an article 5 entity. A statute or licence which delegates coercive cyber-defence, intelligence collection or disruption powers may be different. Governments should not be allowed to obtain a public power's practical effect whilst drafting the instrument so that every operative is described as a commercial supplier.

Article 8 asks about the person or group carrying out the conduct, not about whether software had discretion between human interventions. In *Nicaragua* the Court required effective control of the operation in which the alleged violations occurred; in *Bosnian Genocide* it rejected the broader overall-control standard as a rule of State responsibility, whilst *Tadić* had employed that standard for the distinct question of classifying a conflict.[37] The machine complicates proof of the relevant conduct. If the State instructs "disable Network X" and the agent selects Server Y as a means, the absence of a Server Y instruction should not atomise the operation. If it instructs lawful collection and a contractor's system independently attacks Hospital Z contrary to tested constraints, attribution of the excess is more difficult. The correct level of description—campaign, target class, individual act or consequence—cannot be selected solely to produce attribution or avoid it.

Instructions may be expressed as effects and constraints rather than technical steps. Payment upon a target country's loss of export capacity, supply of exclusive intelligence, approval of civilian-service dependencies, daily receipt of operational telemetry and power to halt are evidence from which instruction, direction or control may be inferred. Payment for a finished public report, no operational access, a genuine prohibition upon external effects and prompt termination upon notice point the other way. No single commercial clause decides the international question; the factual operation does.

Adoption under article 11 is intentionally demanding. In the *Tehran Hostages* case, the initial private seizure and the later State endorsement were treated differently; the ILC commentary accordingly requires acknowledgement and adoption of the conduct as the State's own, not mere approval of a result.[38] A ministry which receives data produced by an autonomous intrusion does not necessarily

[35] International Law Commission, 'Draft Articles on Responsibility of States for Internationally Wrongful Acts, with Commentaries' (2001) UN Doc A/56/10, arts 1–2, 4–11 and commentaries; James Crawford, *State Responsibility: The General Part* (CUP 2013) 113–75.

[36] International Law Commission (n 35) arts 5 and 7; *Noble Ventures Inc v Romania* (Award) ICSID Case No ARB/01/11 (12 October 2005) [69]–[86]; *Jan de Nul NV and Dredging International NV v Egypt* (Award) ICSID Case No ARB/04/13 (6 November 2008) [159]–[173].

[37] *Military and Paramilitary Activities in and against Nicaragua (Nicaragua v United States of America)* (Merits) [1986] ICJ Rep 14 [109]–[115]; *Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia and Herzegovina v Serbia and Montenegro)* [2007] ICJ Rep 43 [398]–[407]; *Prosecutor v Tadić* (Appeals Chamber Judgment) IT-94-1-A (15 July 1999) [116]–[145].

[38] *United States Diplomatic and Consular Staff in Tehran (United States of America v Iran)* (Judgment) [1980] ICJ Rep 3 [58]–[74]; International Law Commission (n 35) art 11, commentary paras 1–10.

adopt every prior act. Its informed direction that collection continue, public identification of the operation as its own, or operational integration of the same hostile system may supply stronger evidence. Domestic law may nevertheless prohibit knowing receipt, use or concealment at a lower threshold without pretending to amend article 11.

Article 16 presents a further boundary. It concerns aid or assistance by one State to another State's internationally wrongful act, with knowledge of the circumstances and the other conditions stated in the article; it is not a general doctrine for assistance to private actors.[39] Where one State supplies a model service, exploit intelligence, credentials or compute to another's operation, the rule may be engaged. Where the immediate recipient is a PMSC, the analysis must first determine its relationship to a State, the applicable primary rules and any direct duties of the assisting State. The ease of saying "AI support" must not abolish these elements.

Designed ignorance operates within, not instead of, those rules. It permits a tribunal to draw ordinary conclusions from a party's deliberate destruction of the evidence by which the elements would be tested. The inference should be strongest where the State had exclusive access to the public objective, selected the contractor, prescribed reporting, could demand the manifest, and chose an arrangement advertised as preserving deniability. It should be weaker where records were genuinely lost to a hostile compromise or never lawfully available to the State. Evidential realism is not a new attribution rule; it is what prevents an old rule from being defeated by architecture.

5.1 A Responsibility Matrix

The same operation may engage several forms of responsibility without contradiction.

The State principal. International responsibility depends upon attribution and breach. Domestic public law examines statutory authority, rationality, human rights, procurement and ministerial accountability. Political responsibility may arise where legal attribution fails but public money and policy created the operation. The State should not be allowed to use the strictness of article 8 as a reason to publish no account to Parliament.

The PMSC company. A company may bear contractual, regulatory, civil and criminal liability under domestic law. It may be designated or sanctioned. International law ordinarily attributes conduct to States rather than making companies direct bearers of the Charter prohibition, although IHL binds parties and individuals in relevant ways and businesses possess responsibilities under instruments including the UN Guiding Principles on Business and Human Rights.[40] The statute should avoid saying that the corporation "committed aggression" unless the relevant criminal law so provides.

Directors and programme managers. Their liability depends upon authorisation, knowledge, control and offence. A director who approves deployment after an evaluation shows indiscriminate selection is differently placed from a non-executive who receives an inaccurate summary. Senior-manager accountability should follow actual decision rights, not title alone.

Developers and integrators. General-purpose development is remote. Integration for a known hostile objective is close. Aiding, encouraging, conspiracy and attempt doctrines supply modes of liability where mental elements are met. Licensing and civil duties may apply at a lower threshold than crime. A statutory presumption should not convert every open-source contributor into an accomplice.

Infrastructure and model providers. Before notice, providers ordinarily supply general services. After credible and specific notice, their reasonable response may include investigation, preservation,

[39] International Law Commission (n 35) art 16 and commentary; Helmut Philipp Aust, *Complicity and the Law of State Responsibility* (CUP 2011) 191–249.

[40] UN Human Rights Council, 'Guiding Principles on Business and Human Rights' (21 March 2011) UN Doc A/HRC/17/31.

account restriction or engagement with authorities, constrained by privacy, contract and human-rights law. A provider integrated into the operation, paid for dedicated capability and informed of target purpose is not equivalent to an unaware host.

The human deployer. This person commonly possesses the final practical authority: connecting credentials, objective and external environment. Deployment is a legally salient act even where later choices are autonomous. Licensing should ensure the deployer is identified.

The model. It bears no punishment. Its configuration and output are evidence; its assets may be seized as infrastructure; its risk may justify restrictions. Creating electronic personality would offer principals a judgment-proof intermediary and should be rejected.

The territorial State. It may have duties after knowledge, capacity and risk arise. Its courts may possess jurisdiction; its authorities may preserve evidence. Hosting one fragment does not make it sponsor. Refusal to act after detailed notice is a different case from inability to see encrypted computation.

The victim organisation. Failure to patch may contribute factually and breach regulatory duties; it does not authorise attack or transfer responsibility from the attacker. The Electoral Commission's security failures and the PRC-linked attribution can both be true. Law should resist a false choice between victim resilience and aggressor responsibility.

The matrix also disciplines remedy. Criminal prosecution addresses persons; damages address loss; procurement sanctions address contractors; asset freezes address economic capacity; countermeasures address State breach; self-defence addresses armed attack. One label—"AI attack"—cannot select among them.

6. Due Diligence and the Territorial Fragment

The *Corfu Channel* judgment stated that every State has an obligation not to allow knowingly its territory to be used for acts contrary to the rights of other States.[41] The precise content and application of due diligence in cyberspace remain contested, including whether a general binding cyber-specific rule exists and what knowledge, capacity and reasonable measures require. Published national positions differ.[42] The principle cannot sensibly demand that a State prevent what it could neither know nor reasonably control.

Distributed agents make the limitation acute. A State may host a short inference task which contains no target identity. Another hosts encrypted memory. A third contains the compromised system from which traffic emerges. Each sees a benign-looking fragment. It is difficult to say that the first State failed to protect external actors where the safeguard necessary to identify the threat would require disproportionate inspection of every computation.

This is where the work of Evelyne Schmid and Ayşe Özge Erceiş is especially valuable. Their analysis of attribution of omissions explains how due-diligence responsibility concerns a State's own failure to take required measures rather than attribution of the private attacker's conduct, and why knowledge, risk, capacity and reasonableness must be kept analytically distinct.[43] The insight should be extended, respectfully, to deliberately fragmented autonomous operations: the hosting State's omission cannot be assessed by pretending that it possessed the victim's assembled picture.

[41] *Corfu Channel (United Kingdom v Albania)* (Merits) [1949] ICJ Rep 4, 22.

[42] UNGA, 'Official Compendium of Voluntary National Contributions on the Subject of How International Law Applies to the Use of Information and Communications Technologies by States' (13 July 2021) UN Doc A/76/136.

[43] Evelyne Schmid and Ayşe Özge Erceiş, 'The Attribution of Omissions: Due Diligence in Cyberspace and State Responsibility' (2023) 33 *Swiss Review of International and European Law* 577.

That analytical separation is indispensable. An omission by State H is attributable to State H because its organs failed to act; the contested questions are whether a primary obligation required action, whether the conditions of that obligation arose, and what measure was reasonably available. It is unnecessary, and usually wrong, to attribute the private intruder to H in order to establish H's responsibility for its own omission. Schmid and Erceiş thereby supply a route to accountability which does not dilute article 8, but the route is exacting: a State cannot fail to prevent a danger of which it neither knew nor ought reasonably to have known, by means it neither possessed nor could lawfully obtain.[44]

States do not yet agree upon one general cyber-specific due-diligence rule. The Netherlands and several other States have described a duty to take reasonable measures where a State knows of wrongful cyber activity from its territory; the United Kingdom has, in published positions, declined to accept that the due-diligence principle has crystallised as a specific rule in cyberspace, whilst still recognising obligations which may require action in particular circumstances.[45] The disagreement must be stated, not wished away. A domestic statute may impose duties upon United Kingdom operators and public authorities; it cannot announce contested custom into existence for every foreign host.

Fragmentation changes each element. Knowledge may exist at the victim but not the host. Risk may be grave when streams are correlated and innocuous when one stream is viewed alone. Capacity may lie with a provider but require judicial authority which the executive cannot dispense with. A feasible measure may be preservation rather than immediate termination, because termination can destroy evidence, affect innocent tenants or displace the operation into a less cooperative State. Reasonableness therefore demands a decision specific to the indicator, service, urgency, safeguards and expected harm.

The controller's position is ordinarily stronger. He chose the objective, routing logic, providers and retention. If he fragmented them for resilience, he can maintain a correlation record; if he fragmented them for deniability, refusal to maintain that record is itself probative. The proposed duty of correlation is consequently imposed upon the orchestrator and prime contractor, not by compelling every territorial State to inspect encrypted computation. It implements the deeper insight in Schmid's work: responsibility for omission should fall where law, knowledge and capacity genuinely meet, not where the last visible packet happened to cross a border.

Notice should be structured accordingly. A bare allegation that "malicious AI" operates somewhere in a provider's region is insufficient. A valid notice should identify the harmful campaign, technical indicators, time window, evidential confidence, requested action, legal basis and foreseeable prejudice to other users. Receipt triggers authentication and assessment, preservation where lawful and urgent, and a reasoned response. Repeated, corroborated notice changes what the host and provider can reasonably deny knowing. The proposed federated system in Papers 3 and 3B is intended to make that progression visible without transferring whole national datasets.

The controller, however, may possess that picture. Regulation should move a duty of correlation towards actors who design or orchestrate high-risk systems. Cloud providers should not inspect content indiscriminately; a PMSC prime contractor and State principal should preserve a manifest of tasks, regions, identities and objective. Where they intentionally distribute the operation to defeat territorial safeguards, a rebuttable evidential inference should follow against them.

---

[44] Schmid and Erceiş (n 26) 586–601; *Corfu Channel (United Kingdom v Albania)* (Merits) [1949] ICJ Rep 4, 22; *Pulp Mills* (n 26) [101], [197].

[45] Government of the Netherlands, 'Letter to the Parliament on the International Legal Order in Cyberspace' (5 July 2019) 4–5; Foreign, Commonwealth & Development Office, 'Application of International Law to States' Conduct in Cyberspace: UK Statement' (3 June 2021); Federal Government of Germany, 'On the Application of International Law in Cyberspace' (March 2021) 3; UNGA, 'Official Compendium of Voluntary National Contributions on the Subject of How International Law Applies to the Use of Information and Communications Technologies by States' (13 July 2021) UN Doc A/76/136.

Host States should be judged upon notice and reasonable capability. Once supplied with authenticated indicators showing that specific infrastructure forms part of an ongoing harmful operation, a State should investigate and take feasible measures consistent with law. Before notice, baseline duties may include abuse contacts, preservation capacity, judicial process and security for government infrastructure. Omniscience is not due diligence.

6.1 The Attraction of Older and Weaker Systems

Security debate tends to imagine a contest between the newest model and the best national defender. An attacker is under no such obligation. He may search for the system whose failure is useful and whose defence is cheapest.

Water and wastewater operators commonly manage long-lived operational technology, remote sites, specialist protocols and equipment which cannot be patched or restarted like an office laptop. Health systems combine clinical devices, identity, scheduling, records, laboratories and suppliers. Insurers and local authorities hold broad personal data and rely upon third-party software. Smaller government agencies may lack continuous monitoring, asset inventories or exercised restoration. The vulnerabilities are organisational as often as novel.[46]

An autonomous system changes four economic variables. It lowers the marginal cost of attempting another target; operates across time zones without fatigue; records and reuses what succeeded; and permits a small human team to supervise several campaigns. It need not be reliable on each attempt if targets are abundant. A five per cent success rate across thousands of poorly maintained services is not comforting.[47]

Automation also compresses response time. A human criminal who discovers access may pause to understand value. An agent configured to pursue an objective can classify, test and act as results arrive. Defenders dependent upon weekly review or office-hours approval face a machine timescale. The proper defensive answer includes rate limits, segmentation, strong identity, allowlisting, offline restoration, monitoring and automated containment; the legal answer includes rapid reporting and authority to coordinate.

The harm is not exhausted by compromise. An agent may make destructive mistakes because it misunderstands a legacy control system. A PMSC that treats unreliability as acceptable where the victim bears the cost engages recklessness. Evaluation must include systems which are fragile, poorly documented and safety-critical, not merely modern web applications upon which benchmarks are convenient.

There is also an equality problem. Large banks can buy teams and redundancy. A rural water authority cannot match a State-financed autonomous campaign. Due diligence which simply orders every victim to “be secure” mistakes resource disparity for fault. National defence must provide shared detection, tested baselines, grants and emergency technical assistance to the weak points upon which stronger systems depend.

---

[46] US Government Accountability Office, *Critical Infrastructure Protection: EPA Urgently Needs a Strategy to Address Cybersecurity Risks to Water and Wastewater Systems* (GAO-24-106744, August 2024); US Government Accountability Office, *Critical Infrastructure Protection: Actions Needed to Address Cybersecurity Challenges Facing the Water Sector* (GAO-26-109159, May 2026); Cybersecurity and Infrastructure Security Agency, ‘CISA Urges Water and Wastewater Systems Sector to Protect OT against Activity Targeting PLCs’ (30 July 2026) https://www.cisa.gov/news-events/alerts/2026/07/30/cisa-urges-water-and-wastewater-systems-sector-protect-ot-against-activity-targeting-plcs accessed 5 August 2026; US Environmental Protection Agency, ‘Enforcement Alert: Drinking Water Systems to Address Cybersecurity Vulnerabilities’ (May 2024).

[47] National Cyber Security Centre, *Annual Review 2025* (14 October 2025) 34–37; Cybersecurity and Infrastructure Security Agency, *Cross-Sector Cybersecurity Performance Goals* (updated 2023); National Institute of Standards and Technology, *Cybersecurity Framework (CSF) 2.0* (NIST CSWP 29, February 2024); Network and Information Systems Regulations 2018, SI 2018/506.

7. Use of Force, Armed Attack and the Autonomous Choice of Target

A cyber operation may be a prohibited use of force or armed attack according to scale and effects. The United Kingdom's position expressly adopts that functional approach.[48] Autonomy does not alter the threshold. It alters how intent, target selection, aggregation and imminence are proved.

There is no sound reason to reserve the legal gravity of an armed attack to an object which moves by gunpowder. A rifle, a missile and a keyboard differ in mechanism; the Charter's concern is the coercive force and consequence for which the mechanism is employed. An operation which manipulates industrial control so as to kill, destroy or disable upon a scale comparable to conventional attack is not rendered peaceful by the absence of a projectile. Published positions of the United Kingdom, United States, France, Germany, Australia, Canada and others accept, in materially different formulations, that cyber operations may reach the use-of-force and armed-attack thresholds according to their scale and effects.[49] The proposition is now orthodox enough to state firmly, though the classification of any particular campaign remains dependent upon fact.

Four terms must nevertheless be kept separate. **Kinetic effect** ordinarily describes physical movement, damage or injury. **Use of force** is the article 2(4) threshold. **Armed attack** is the graver article 51 threshold upon which self-defence depends. An **attack** under IHL is an act of violence against the adversary, whether in offence or defence, and operates within an existing armed conflict.[50] One cyber operation may satisfy several categories; it need not. A keyboard can be the means of an armed attack without every unlawful access becoming "kinetic", and an IHL attack need not determine the Charter question.

Where data are deleted without immediate physical destruction, the classification is harder and the stakes are not smaller. Erasure of patient records, loss of drug-allergy information, interruption of emergency dispatch, manipulation of energy distribution in winter, or corruption of water treatment may cause death through the service which fails rather than through the computer which breaks. Legal analysis should count those causal effects. The National Audit Office's account of WannaCry recorded cancelled appointments, diverted ambulances and widespread service disruption, notwithstanding that the malware was not designed as a State armed attack.[51] The example proves consequence, not attribution.

---

[48] Foreign, Commonwealth & Development Office, 'Application of International Law to States' Conduct in Cyberspace: UK Statement' (3 June 2021) https://www.gov.uk/government/publications/application-of-international-law-to-states-conduct-in-cyberspace-uk-statement/application-of-international-law-to-states-conduct-in-cyberspace-uk-statement accessed 31 July 2026.

[49] Foreign, Commonwealth & Development Office (n 45); United States, 'International Law and Stability in Cyberspace' (2016) 50 *Digest of United States Practice in International Law* 821; France, Ministry of the Armies, *International Law Applied to Operations in Cyberspace* (2019) 7; Federal Government of Germany (n 45) 5–6; Australian Government, *Australia's Position on How International Law Applies to State Conduct in Cyberspace* (2020); Government of Canada, *International Law Applicable in Cyberspace* (22 April 2022) paras 43–52; New Zealand Ministry of Foreign Affairs and Trade, *The Application of International Law to State Activity in Cyberspace* (1 December 2020) paras 14–18.

[50] Charter of the United Nations arts 2(4) and 51; Protocol I (n 27) art 49; Michael N Schmitt (ed), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (CUP 2017) rr 69, 71 and 92; Heather Harrison Dinniss, *Cyber Warfare and the Laws of War* (CUP 2012) 74–99; ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts* (October 2024) ch 5.

[51] Comptroller and Auditor General, *Investigation: WannaCry Cyber Attack and the NHS* (HC 2017–19, 414); Comptroller and Auditor General, *Digital Transformation in the NHS* (HC 2019–21, 317) paras 1.5 and 2.13–2.15; Information Commissioner's Office, 'London Borough of Hackney Reprimanded Following Cyber-attack' (17 July 2024) https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2024/07/london-borough-of-hackney-reprimanded-following-cyber-attack/ accessed 5 August 2026; Information Commissioner's Office, *Advanced Computer Software Group Limited: Monetary Penalty Notice* (26 March 2025).

Risk must also be judged before the casualty list is complete. For regulation and defensive preparation, the relevant measure is the **maximum reasonably credible and foreseeable harm** of the authorised operation: harm supported by target, access, system dependency and plausible failure mode. It is not the most imaginative catastrophe which counsel can invent after an innocuous probe. For classification after an event, actual and proximate effects, intended effects, scale, duration, reversibility, target, military character and context all bear weight. The two enquiries should not be confused merely because both use the word “harm”.

Autonomy makes aggregation decisive. A controller may instruct one system to maximise pressure and allow it to distribute modest actions across a hundred local services. If common objective, operation and control are proved, the law should examine the combined campaign rather than accept the machine’s task boundaries as legal boundaries. This principle is no licence to aggregate unrelated events. Shared malware, a common model or temporal proximity may be indicators; the evidential chain must establish the joined operation.

Where an agent attacks several water stations, each local effect may be modest and the combined effect grave. If a common objective and controller are proved, the campaign should be assessed as a whole. The machine’s distribution of tasks cannot force the victim to treat each as legally unrelated. Conversely, similar attacks by unrelated systems cannot be aggregated merely because they use the same model or language.

Intent is not invariably an element of the prohibition of force as such, but it matters to classification, response and individual liability. A deployer who authorises “maximum disruption” across critical infrastructure cannot plausibly treat the system’s selection of a hospital as wholly accidental where hospitals were not excluded and were technically visible. A system escaping a bounded test environment raises a different case; negligence, product safety, State due diligence and cessation may apply without a hostile armed attack by the developer’s State.

Imminence becomes harder where an agent is persistent. The State may detect reconnaissance, tool staging and autonomous evaluation without knowing the final target. The “last window of opportunity” cannot be equated to the first suspicious packet. Evidence should include objective, capability, proximity to protected systems, persistence, previous attempts and whether human approval remains before harmful effect. A broad permission to act wherever a model might later choose would erase article 2(4).

Response against infrastructure also requires distinction. A commercial data centre is not a military objective because malicious code once passed through it. A particular server, virtual workload or account may effectively contribute to military action; neutralisation must offer a definite military advantage and precautions must minimise effects upon civilian tenants.[52] Precise cyber or law-enforcement action may be preferable to kinetic strike. Where lethal force is considered, the legal and evidential bar must remain high.

That last proposition does not require that direct force be removed from consideration. If an autonomous operation amounts to an armed attack, article 51 may permit necessary and proportionate self-defence against the responsible actor, subject to attribution, sovereignty, reporting and the law governing the means used.[53] A person or small private group can in principle conduct an armed attack; the victim is not obliged to pretend that State insignia are a physical property of force. Yet a known server location does not prove that every occupant, building or territorial State is the attacker. The

[52] Protocol I (n 27) arts 48, 51, 52 and 57.

[53] Charter of the United Nations art 51; *Oil Platforms (Islamic Republic of Iran v United States of America)* (Judgment) [2003] ICJ Rep 161 [51], [73]–[77]; *Armed Activities on the Territory of the Congo (Democratic Republic of the Congo v Uganda)* (Judgment) [2005] ICJ Rep 168 [146]–[147]; *Legal Consequences of the Construction of a Wall in the Occupied Palestinian Territory* (Advisory Opinion) [2004] ICJ Rep 136 [139].

speed demanded by defence strengthens the case for pre-authorised evidential standards and precise capabilities; it does not make uncertainty disappear.

Proportionality in self-defence concerns the defensive response necessary to halt or prevent the armed attack, not a tariff of equal suffering. Expected future harm may matter where continuation or recurrence is evidenced. Punishment for what might conceivably have happened is a different thing. The statutory regime should therefore record the system's objective, capacity, observed trajectory and termination state in real time, so that Ministers need not choose between impotence and conjecture when minutes matter.

8. International Humanitarian and Criminal Law

If an armed conflict exists and the autonomous operation has the required nexus, IHL governs conduct of hostilities. The party must distinguish civilians and civilian objects, apply proportionality in attack and take feasible precautions. Delegating target identification does not delegate the obligation. A system used to select targets must be capable, within its context, of supporting those judgments or be confined to recommendations subject to competent human assessment.

The word "human" alone is insufficient. An operator confronted with hundreds of machine recommendations and seconds to approve may be a ceremonial rubber stamp. Meaningful control requires information, time, competence, authority to refuse and a system which communicates uncertainty and the basis of classification. Where these are absent, the human does not repair automation; the procedure disguises it.

War crimes are committed by natural persons. The Rome Statute requires the material and mental elements of the offence, subject to its jurisdiction and admissibility rules.[54] A commander or superior may incur responsibility under article 28 where its conditions are met. A programmer far from deployment is not automatically a perpetrator; liability depends upon contribution, knowledge, intent and the mode of responsibility. The PMSC commander who knowingly deploys an indiscriminate system is differently situated from a library developer unaware of hostile use.

Article 30's default mental element of intent and knowledge does not translate easily to probabilistic choice. A person may intend to cause a consequence where he means to cause it or is aware it will occur in the ordinary course of events. Testing, target distributions, known false-positive rates, prior incidents and ignored warnings become important. "The model did it" is not an answer where the accused deployed it because its ordinary operation produced the prohibited class of consequence.

Domestic law can reach negligence and recklessness which international criminal law does not. The Computer Misuse Act 1990 criminalises unauthorised access and impairment with specified mental elements; conspiracy, attempt, assisting or encouraging, fraud, data and national-security offences may apply.[55] Corporate attribution and failure-to-prevent models should be examined for PMSC deployment. Criminal breadth must be matched by security-research defences and public-interest safeguards so that defensive testing is not driven away.

Article 36 of Additional Protocol I requires a party, in the study, development, acquisition or adoption of a new weapon, means or method of warfare, to determine whether its employment would in some or all circumstances be prohibited.[56] An autonomous cyber capability may be a weapon, means or method

[54] Rome Statute of the International Criminal Court (adopted 17 July 1998, entered into force 1 July 2002) 2187 UNTS 3, arts 25, 28 and 30.

[55] Computer Misuse Act 1990, ss 1–3ZA.

[56] Protocol I (n 27) art 36; ICRC, *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare* (January 2006); Ministry of Defence, *Defence Artificial Intelligence Strategy* (15 June 2022) 22–25.

according to its design and intended use; a State should not evade review by procuring a continuing service rather than purchasing a named artefact. Review must extend to the integrated configuration which receives objective, target data, tools and authority. Approval of a base model proves little where the contractor later changes scaffolding, permissions and reward.

The review should ask what the system can sense or infer about protected status, how uncertainty is expressed, what operational environment was tested, whether effects can be limited, how target and route drift are detected, and whether termination works after loss of the principal node. It should identify uses which are lawful only within a constrained environment and uses which cannot be made lawful because their effects cannot be directed or limited. The ICRC's position upon autonomous weapons emphasises that the operator may not know the specific target, time or place selected after activation; that description maps closely, though not identically, upon an autonomous cyber agent choosing an address and moment.[57]

IHL responsibility remains with the party and the natural persons who act for it. A target-selection system cannot itself perform the commander's legal judgment in a manner which relieves the commander. It may process information and recommend; the decision-maker must possess reasonable grounds, consider feasible precautions and cancel or suspend where it becomes apparent that the object is protected or expected incidental harm excessive. The law does not demand metaphysical certainty. It demands a process adequate to the decision whose consequences the party chooses to impose.

Individual criminal responsibility becomes difficult where contribution is modular. The Rome Statute distinguishes commission, ordering, soliciting, inducing, aiding or abetting, common-purpose contribution and superior responsibility; each has its own act and mental requirements.[58] The ICTY's jurisprudence upon assistance, purpose and remoteness, and the ICC's judgments upon contribution and command, warn against a single elastic category of "involvement".[59] A programmer supplying general memory code is not placed with a PMSC officer who integrates that code into a system whose evaluations display indiscriminate harm; the State official who knowingly orders the prohibited objective is differently placed again.

The time-zone chain does not preclude common purpose, but it makes proof granular. One participant may know the strategic object but not the criminal means; another may know the target class but not the State principal; a third may know that logs are deleted in order to obstruct investigation. Communications, specifications, payment, access, test results and later concealment may establish knowledge or contribution. The fact that no individual wrote the whole system is no more a defence than the fact that no individual manufactured a whole missile.

Command and superior responsibility require effective command and control, or effective authority and control, and the other conditions of the governing provision. Article 28 is not satisfied by job title or by the abstract ability to influence corporate policy.[60] In an agentic operation, the relevant powers may be divided: a programme director controls personnel, a technical officer controls credentials, a State liaison controls objectives, and a provider can terminate the account. The law should record those powers

[57] ICRC, *ICRC Position on Autonomous Weapon Systems* (n 23) 5–12; ICRC, *Autonomous Weapon Systems and International Humanitarian Law: Selected Issues* (November 2025); CCW Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, 'Report of the 2023 Session' (24 May 2023) UN Doc CCW/GGE.1/2023/2.

[58] Rome Statute (n 54) arts 25, 28 and 30; *Prosecutor v Bemba Gombo* (Judgment on the Appeal) ICC-01/05-01/08-3636-Red (8 June 2018); *Prosecutor v Ntaganda* (Judgment) ICC-01/04-02/06-2359 (8 July 2019); *Prosecutor v Lubanga Dyilo* (Judgment) ICC-01/04-01/06-2842 (14 March 2012).

[59] *Prosecutor v Furundžija* (Trial Judgment) IT-95-17/1-T (10 December 1998) [190]–[249]; *Prosecutor v Perišić* (Appeals Judgment) IT-04-81-A (28 February 2013); *Prosecutor v Šainović and others* (Appeals Judgment) IT-05-87-A (23 January 2014) [1649]–[1651].

[60] Rome Statute (n 54) art 28; *Bemba* (n 58) [167]–[194]; *Prosecutor v Halilović* (Trial Judgment) IT-01-48-T (16 November 2005) [54]–[100].

before harm. It should not invent one commander after the event because an organisational chart is simpler than the facts.

Domestic criminal law supplies additional, but not unlimited, reach. The Serious Crime Act 2007 addresses intentional encouragement or assistance; conspiracy requires the agreement and relevant fault; *Jogee* confirms that foresight is evidence from which intent may be inferred, not a substitute for the intent required by accessorial liability.[61] Recklessness under *R v G* asks whether the defendant was aware of a risk and, in the circumstances known to him, it was unreasonable to take it.[62] These principles are capable of addressing a deployer who knows that the system may escape scope and proceeds; they do not criminalise every supplier merely because misuse was conceivable.

Corporate responsibility requires similar precision. The identification doctrine, statutory senior-manager rules and specific failure-to-prevent offences operate differently.[63] A new autonomous-strategic-systems offence should state whose knowledge and decision are attributed to the company, what reasonable prevention procedures mean, and how an ordinary provider or researcher is excluded. It should not rely upon the hope that a Victorian concept of the directing mind will happen to locate a distributed technical decision.

9. The Autonomous System as Third Party and as Evidence

The term “third party” has two meanings. Institutionally, the system stands between State principal and victim, appearing to make operational choices. Evidentially, it generates records capable of showing what people decided: objective files, tool permissions, model versions, task graphs, approvals, budgets, route policies and stop conditions.

Law should exploit the second meaning. A statutory attribution manifest should be mandatory for government and PMSC high-risk operations. It should record:

- the authorised objective and prohibited objectives;
- responsible State and contractor officials;
- model, framework and material version;
- tool and credential permissions;
- target and geographical constraints;
- data sources and validation;
- human approval points;
- compute regions and providers;
- persistence and recovery rules;
- suspension and revocation mechanisms;
- audit-log locations and retention;
- subcontractors and module provenance.

The manifest need not be public during operations. It must be available to authorised inspectors, courts and parliamentary oversight. Technical telemetry should be cryptographically protected against alteration, with separate access controls for intelligence and personal data. Absence or destruction should have legal consequence.

---

[61] Serious Crime Act 2007, ss 44–46; Criminal Law Act 1977, s 1; *R v Jogee; Ruddock v The Queen* [2016] UKSC 8, [2017] AC 387 [66]–[87]; *R v Saik* [2006] UKHL 18, [2007] 1 AC 18 [14]–[28].

[62] *R v G* [2003] UKHL 50, [2004] 1 AC 1034 [41]; *R v Cunningham* [1957] 2 QB 396 (CA) 399.

[63] *Tesco Supermarkets Ltd v Nattrass* [1972] AC 153 (HL); *Meridian Global Funds Management Asia Ltd v Securities Commission* [1995] 2 AC 500 (PC); Economic Crime and Corporate Transparency Act 2023, ss 196 and 199–206; Corporate Manslaughter and Corporate Homicide Act 2007, ss 1–2.

A rebuttable presumption is justified where a State contractor was under a duty to keep the manifest, the evidence lies peculiarly within its control, serious harm occurred within the authorised capability and it fails without reasonable explanation to produce records. The presumption should concern deployment, control and knowledge facts; it cannot reverse the criminal burden upon the ultimate offence contrary to fair-trial rights.[64]

Model output should not be treated as an oracle. Natural-language "reasoning" may be post hoc or unreliable. Greater weight belongs to objective configuration, actual tool calls, system state, external logs and human messages. Investigators need reproducibility where feasible, but stochastic systems may not repeat. Preservation of the actual run is therefore more valuable than later demonstrations.

9.1 Causation, Foreseeability and the Alleged Intervening Machine

The system will often be described as a *novus actus*: an independent choice breaking the chain between human and harm. The argument should ordinarily fail where the very reason for deployment was that the system would choose adaptive means within an authorised field. An intervening act breaks legal causation only under the law governing the particular claim; autonomy supplies no universal answer.

Three counterfactuals assist. First, had the human removed the objective, would the harmful operation have occurred? Secondly, had he withheld credentials, tools or external access, could it have occurred through this system? Thirdly, after warning signs, could he reasonably have stopped or confined it? Affirmative answers do not by themselves prove an offence, but they locate causal power.

Foreseeability should operate at the correct level. It may be unforeseeable that an agent will choose Hospital X at 03:17 through a particular provider. It may be entirely foreseeable that a system authorised to maximise economic interruption without protected-sector exclusions will select a hospital supplier. Insisting upon prediction of the exact path would make delegated discretion an immunity.

The converse protects legitimate developers. A general language model distributed for research may be altered, connected to stolen credentials and used in a hostile system years later. "AI can be misused" is too general to make the original developer the legal cause of every act. Proximity increases through customisation, integration, knowledge, control, profit from the operation and ability to prevent it.

Product liability concepts may address defective safety claims, but hostile use is not always defect. A model may perform exactly as its malicious operator intends. The regulator should therefore distinguish **unsafe product**, **unsafe deployment** and **intentional weaponisation**. The manufacturer may own the first; the operator the second; principal and operator the third. Responsibility may overlap where the manufacturer knowingly supplies a tailored capability.

Foreseeable misuse duties must be proportionate. Providers can monitor abuse indicators and enforce terms; they cannot read every private computation without destroying security and privacy. Higher duties are justified for a dedicated PMSC service, unusual offensive tool access, repeated warnings or State contract. General consumer access supports general safeguards, not universal surveillance.

Loss allocation follows causation but should not await perfect attribution. A victim-compensation fund financed by licensed operators and public procurement fees can pay urgent recovery, preserving later recourse against responsible actors. Otherwise the evidential complexity engineered by the operation becomes an additional weapon: the victim bears loss for years while principals litigate which module chose it.

[64] *Sheldrake v DPP; Attorney General's Reference (No 4 of 2002)* [2004] UKHL 43, [2005] 1 AC 264; *R v Lambert* [2001] UKHL 37, [2002] 2 AC 545; *Salabiaku v France* (1988) 13 EHRR 379 [28]; *Big Brother Watch v United Kingdom* (2022) 74 EHRR 17 [332]–[347].

10. A Statutory Model for the United Kingdom

The proposed UK Economic Warfare Prevention and Response Act should contain a Part on autonomous strategic systems.

The Part is necessary because the present law is distributed according to the object harmed, the access obtained, the data processed, the public power exercised and the person involved. The Computer Misuse Act reaches unauthorised access and impairment; data-protection law imposes security and breach duties; the Network and Information Systems Regulations regulate specified essential and digital services; the National Security Act addresses forms of foreign-power threat activity; procurement law governs the public contract. None creates an operation-wide duty to identify who delegated coercive decision authority to an autonomous system, where it may act and how it is stopped.[65]

Designated operation. A "designated autonomous strategic operation" should mean deployment of a system with material delegated discretion to select a person, object, technique, jurisdiction or continuation of action capable of unauthorised access, impairment, coercive economic effect or contribution to hostilities. Research in controlled environments, ordinary defensive administration and systems without external action should be excluded or separately regulated.

Licensing. A United Kingdom person should require authority to provide or operate such a system for a foreign State, armed group, PMSC or conflict party. The licence should identify objective classes, territories, providers, safeguards and reporting. An emergency defensive licence may be granted rapidly.

Non-delegable accountability. Government departments and prime PMSCs should appoint a responsible officer. Subcontracting should not remove the duty to maintain the manifest, test constraints, monitor operation and retain termination capacity.

Evaluation. Before deployment, the operator should test against protected-system scenarios, false target classifications, route leakage, prompt or data manipulation, persistence after stop, credential overreach and log failure. The test should establish limits, not certify universal safety.

Change control. Material changes to model, tools, permissions, objective or target class should require recorded approval. Continuous learning upon live hostile data should be prohibited unless specifically authorised and bounded; otherwise the evaluated system is not the deployed system.

Stop and recovery. A secure revocation mechanism should exist across every authorised node. Recovery should distinguish infrastructure failure from a lawful stop. A system which cannot reliably stop should not receive authority to act outside a controlled range.

Incident reporting. Loss of control, action outside scope, protected-target contact, unauthorised replication, provider compromise and material log loss should be reported immediately. A provider discovering misuse should preserve evidence and notify through a protected channel, subject to human-rights and data safeguards.

Designed fragmentation. Deliberately distributing code, data or compute to conceal a prohibited operation or defeat a lawful investigation should be an aggravating factor and, where accompanied by the relevant intent, an offence. Ordinary resilience, privacy engineering and multinational cloud use must not be criminalised.

[65] Computer Misuse Act 1990, ss 1–3ZA; Data Protection Act 2018, ss 66 and 149–159; UK GDPR, arts 5(1)(f), 32–34; Network and Information Systems Regulations 2018, SI 2018/506, regs 10–12; National Security Act 2023, pts 1–2; Procurement Act 2023; Investigatory Powers Act 2016, pts 5–6.

Procurement transparency. Departments should report classified details to Parliament's Intelligence and Security Committee or another competent body, and publish aggregate use, incidents and contractors. Commercial confidentiality cannot wholly displace constitutional accountability for State cyber operations.

Civil remedy. Victims should possess a cause of action against United Kingdom operators who intentionally or recklessly deploy outside authority, subject to carefully confined national-security procedure. Compensation does not replace public accountability; it gives error a cost borne by the controller rather than only by the attacked organisation.

10.1 Regulation of Capability without Publishing a Blueprint

Oversight bodies need sufficient technical understanding to test an operation without placing operational methods in public legislation. The statute should therefore require **assurance cases**. The operator must state the capability claimed, environment assumed, hazards identified, controls adopted, evidence supporting each control, residual risk and authority accepting it. The regulator challenges the case; it does not prescribe one model architecture.[66]

Red-team evaluation should occur in controlled cyber ranges representative of likely protected systems. Results should record not only task success but unsafe scope expansion, false classification, inability to stop, logging gaps, use of unauthorised routes and behaviour under misleading data. A system which fails to complete the hostile task may still be dangerous if it damages unrelated services.

Tool permissions should follow least privilege. The planner need not hold execution credentials; the reporting agent need not modify systems; the verifier should be technically capable of halting rather than merely advising. Privileged actions should use short-lived, purpose-bound credentials and be attributable to a particular operation. These are defensive controls, not operational instructions.

Network and compute boundaries should be explicit. Approved regions, providers and egress destinations should be allowlisted. The operator should know when a task leaves them. Where lawful covert operations require concealment, the exception should be authorised and logged by a public official rather than embedded as a generic PMSC discretion.

Data governance is equally important. Target intelligence may be wrong, manipulated or stale. The system should distinguish source confidence and avoid treating model-generated inference as verified fact. Protected categories—medical, humanitarian, diplomatic, electoral—should trigger heightened approval. The objective is not a universal ethical filter; it is a legal constraint mapped to the operation.

Independent evaluation must include supply chain. Open-weight models, fine-tunes, agent libraries and container images may change. Reproducible builds, signed artifacts and version pinning make it possible to identify what ran. A software bill of materials records components; the authority bill described above records external powers and decision owners. Both are necessary.

Finally, the regime should contain a research safe harbour. Bona fide vulnerability research in authorised environments, publication of general defensive knowledge and creation of dual-use libraries should not become criminal merely because a hostile actor later misuses them. The safe harbour should require good faith, proportionate handling and no intentional facilitation of a specified hostile operation. Strong security depends upon research; indiscriminate liability would leave the field to States and criminals.

[66] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1, January 2023); National Institute of Standards and Technology, *Generative AI Profile* (n 9); Department for Science, Innovation and Technology, *AI Cyber Security Code of Practice* (January 2025); National Cyber Security Centre, *Guidelines for Secure AI System Development* (n 9).

10.2 International Cooperation and Evidence

Autonomous operations cross borders faster than mutual legal assistance. The Second Additional Protocol to the Budapest Convention provides mechanisms including direct cooperation with service providers and emergency assistance, subject to its terms and safeguards.[67] The European Union's electronic-evidence regime creates production and preservation mechanisms within its legal order. The United Kingdom should seek interoperable, rights-compliant procedures rather than one global database.

Requests should identify operation, legal basis, data category, urgency and retention. Providers should preserve before content is transferred. Technical indicators may be shared rapidly; subscriber identity and content require stronger process. Emergency channels should be exercised before crisis, with authentication preventing attackers from issuing fraudulent preservation or shutdown requests.

States should recognise a standard attribution manifest for publicly procured autonomous operations. No State is likely to disclose offensive objectives to every foreign host. It can preserve a sealed manifest with independent custody, capable of later production to an agreed tribunal, inspector or parliamentary body. A refusal to create any record should carry diplomatic and evidential weight.

International rules should also address termination. Where a State notifies another that specified infrastructure is part of a continuing autonomous operation, the host should acknowledge, investigate and report feasible action. The requesting State must supply enough evidence and respect sovereignty; the host must not use demands for impossible proof as permanent shelter. Federated correlation, developed in Papers 3 and 3B, allows each to test indicators without disclosing whole national datasets.

Cooperation cannot make all operations attributable. It can remove the procedural advantage gained from speed and fragmentation. The aim is not universal surveillance. It is that a government building an autonomous proxy should no longer be able to rely upon the time required for five lawful requests exceeding the time required for one machine campaign.

11. Objections

It will be said that manifest and logging duties create a target for espionage. They do. Fragmentation of sensitive records, encryption and role-based access are necessary. The alternative is a State operation for which no accountable official can later reconstruct the decision. Security cannot mean deliberate constitutional amnesia.

It will be said that adversaries will ignore the rules. Some will. Domestic law governs United Kingdom procurement, companies, citizens, infrastructure and cooperation; it also supplies the standard by which evidence is requested from allies and sanctions imposed upon foreign actors. The fact that murderers ignore homicide law has never been a complete argument against defining the offence.

It will be said that models change too quickly for legislation. Parliament should fix decision rights, duties and legal thresholds; technical schedules and codes may change. "AI" need not be defined by architecture. Material delegated discretion and capacity for external effect are durable concepts.

It will be said that mandatory human approval makes defence too slow. The model does not require approval for every benign defensive act. Approval points should follow consequence: isolation of one

[67] Second Additional Protocol to the Convention on Cybercrime on Enhanced Co-operation and Disclosure of Electronic Evidence (opened for signature 12 May 2022) CETS No 224; Regulation (EU) 2023/1543 of the European Parliament and of the Council of 12 July 2023 on European Production Orders and European Preservation Orders for electronic evidence in criminal proceedings [2023] OJ L191/118; Directive (EU) 2023/1544 of the European Parliament and of the Council of 12 July 2023 laying down harmonised rules on designated establishments and legal representatives for gathering electronic evidence in criminal proceedings [2023] OJ L191/181.

owned endpoint differs from destructive action against a foreign hospital. Speed is designed through prior authority, bounded playbooks and recorded emergency discretion.

It will be said that this paper overstates current capability. Current systems fail often, require scaffolding and remain inferior to experts upon many open-ended tasks. The paper states those limitations. It also states the fact which caution sometimes conceals: an attacker chooses targets. He need not defeat the best-defended ministry if a poorly secured water company, insurer or local agency supplies adequate harm. Advancement measured in months should be regulated before reliability reaches the level at which every sceptic is personally convinced.

It will be said that a rebuttable presumption punishes missing data where cyber operations inevitably lose logs. The presumption should arise only after a prior statutory duty, serious harm, peculiar control and absence without reasonable explanation. A provider proving compromise, technical failure or lawful destruction may rebut it. Deliberate opacity should not enjoy the same evidential position as unavoidable loss.

It will be said that the United Kingdom cannot regulate foreign PMSCs. It can regulate public contracts, United Kingdom persons, designated services, financial access, licences and cooperation. Extra-territorial criminal jurisdiction requires a genuine nexus and fair process. The proposal does not claim universal legislative power; it makes domestic capability harder to rent for unaccountable harm and supplies conditions for market access.

The deepest objection is that responsibility without final human target selection punishes intention which never existed. That confuses intention to select a named object with intention to deploy a selection process. A person may lawfully authorise a system to identify hostile military radar within strict confidence and review conditions; an error need not be intentional. A person who authorises a system to locate any vulnerable organisation capable of causing economic pain has intentionally selected vulnerability and pain as the rule, even if he does not know the name returned. Criminal guilt still requires the offence’s mental element. Regulatory responsibility for deployment can arise earlier because society need not wait for proof of a war crime before requiring logs and control.

There is also a democratic objection in the opposite direction: public disclosure of State AI operations may reveal sources and methods. The answer is tiered accountability. The responsible Minister, Attorney General, judicial commissioner and parliamentary committee may receive detail; the public receives legal basis, aggregate use, contractors, incidents and remedial action. Secret evidence may sometimes be necessary. Secret competence without any external institution capable of checking it is not.

Finally, some will prefer an international treaty before domestic action. Negotiation should begin, but it will move more slowly than capability. The United Kingdom already controls procurement, licences, offences and contractors. A domestic manifest standard can be adopted by allies and incorporated into contracts tomorrow; it can later inform custom or treaty. Waiting for universal agreement would award the first years of autonomous proxy warfare to whichever actor values opacity most.

12. Conclusion

The autonomous third party is not a person born inside a computer. It is a governance arrangement: human principals distribute decisions amongst contractors, developers, models and infrastructure until the final act appears to belong to none of them. States already use private actors to obscure responsibility. AI permits objective, target, means, time and route to be separated again.

Present capability is enough to matter. Agents reproduce vulnerabilities, operate across experimental networks and have reportedly executed substantial stages of actual campaigns. Open-weight systems reduce dependence upon provider safeguards. Old and weak systems provide targets without any new technical invention. The claim that autonomous attack is "fairyland" survives only by comparing today's system with an omnipotent machine rather than with the vulnerable organisations it may be tasked to find.

Law should respond by following allocated authority. The State official who defines the objective, the PMSC which accepts it, the developer who supplies a material hostile module, the operator who deploys, the provider upon notice, and the person who can stop are not identically responsible; none should disappear because the model selected an IP address. International attribution remains governed by organs, delegated authority, instructions, direction, control and adoption. Domestic law should add procurement, licensing, manifests, evidential presumptions, corporate duties and remedies.

The decisive protection is a complete record held before harm: who wanted what, who allowed which means, where it could act, what it was forbidden to touch, who could stop it and who chose not to. If a State or PMSC builds the chain so that no one knows, the ignorance is part of the design. Responsibility should begin there.

This is not a proposal to imprison a programmer because code travelled farther than expected, nor to make a cloud provider answer for every encrypted tenant. It is a refusal to let the most informed actors purchase ignorance from the least informed. The prime contractor knows that modules form one system; the deployer knows that the system has external authority; the State knows the effect for which it pays. Duties should sit where information and control can be assembled, with lighter notice-based obligations at the general provider and ordinary researcher.

The autonomous system may eventually exceed the architectures described here. The statute does not depend upon their permanence. Objective, authority, external effect, constraint, continuation and control remain intelligible even when models improve. A law built upon model names will age in months; a law built upon delegated coercive decision will follow the power.

The State which wishes to use such power should not be rendered impotent. It should be made competent: able to act quickly, to stop what it starts, to prove what it authorised, to distinguish protected systems, and to defend its conclusion before institutions entitled to ask. The deterrent value of capability is increased, not diminished, when adversaries know that covert delegation will not erase the chain leading back to its purchaser.

There is no virtue in waiting until an agent is infallible. Warfare has always used unreliable men, machines and intelligence; their unreliability enlarges the need for command. The present system can be cheap, persistent and sufficiently successful against weak targets. That combination is already a public threat. Law which demands perfection before recognising agency will discover autonomy only after the first preventable national failure.

The moment for responsibility is deployment, whilst purpose, permission and control may still be proved, preserved and subjected to law.

Bibliography

Books

Aust HP, *Complicity and the Law of State Responsibility* (CUP 2011)

Crawford J, *State Responsibility: The General Part* (CUP 2013)

Dinniss HH, *Cyber Warfare and the Laws of War* (CUP 2012)

Schmitt MN (ed), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (CUP 2017)

Articles and research papers

Buchan R, 'Cyberspace, Non-State Actors and the Obligation to Prevent Transboundary Harm' (2016) 21 *Journal of Conflict and Security Law* 429

Crootof R, 'War Torts: Accountability for Autonomous Weapons' (2016) 164 *University of Pennsylvania Law Review* 1347

Fang R and others, 'LLM Agents Can Autonomously Exploit One-day Vulnerabilities' (2024) arXiv:2404.08144

Fang R and others, 'Teams of LLM Agents Can Exploit Zero-day Vulnerabilities' (2024) arXiv:2406.01637

Kinniment M and others, 'Measuring AI Ability to Complete Long Tasks' (2025) arXiv:2503.14499

Park JS and others, 'Generative Agents: Interactive Simulacra of Human Behavior' (2023) 36 *Proceedings of the ACM Symposium on User Interface Software and Technology* art 2

Schick T and others, 'Toolformer: Language Models Can Teach Themselves to Use Tools' (NeurIPS 2023) 68539

Schmid E and Erceiş AÖ, 'The Attribution of Omissions: Due Diligence in Cyberspace and State Responsibility' (2023) 33 *Swiss Review of International and European Law* 577

Shao H and others, 'An Empirical Study on Large Language Models in Real-world Cybersecurity Tasks' (2024) arXiv:2403.18624

Shinn N and others, 'Reflexion: Language Agents with Verbal Reinforcement Learning' (NeurIPS 2023) 8634

Wu Q and others, 'AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation' (2023) arXiv:2308.08155

Yao S and others, 'ReAct: Synergizing Reasoning and Acting in Language Models' (ICLR 2023)

Zhu Y and others, 'CVE-Bench: A Benchmark for AI Agents' Ability to Exploit Real-World Web Application Vulnerabilities' (2025) arXiv:2503.17332

Government and institutional materials

AI Security Institute, 'Cyber & Autonomous Systems' (2026) https://www.aisi.gov.uk/category/cyber accessed 5 August 2026

—— 'How Fast Is Autonomous AI Cyber Capability Advancing?' (13 May 2026) https://www.aisi.gov.uk/blog/how-fast-is-autonomous-ai-cyber-capability-advancing accessed 5 August 2026

—— 'Incident Report: Unsanctioned Agent Behaviour during Cyber Testing' (5 August 2026) https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing accessed 5 August 2026

—— 'Measuring AI Agents' Progress on Multi-Step Cyber Attack Scenarios' (16 March 2026) https://www.aisi.gov.uk/research/measuring-ai-agents-progress-on-multi-step-cyber-attack-scenarios accessed 5 August 2026

—— 'Our Evaluation of Claude Mythos Preview's Cyber Capabilities' (13 April 2026) https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities accessed 5 August 2026

—— ‘Our Evaluation of OpenAI’s GPT-5.5 Cyber Capabilities’ (30 April 2026) https://www.aisi.gov.uk/blog/our-evaluation-of-openais-gpt-5-5-cyber-capabilities accessed 5 August 2026

—— *Frontier AI Trends Report* (2026) https://www.aisi.gov.uk/frontier-ai-trends-report accessed 5 August 2026

Anthropic, ‘AI Models on Realistic Cyber Ranges’ (16 January 2026) https://www.anthropic.com/research/cyber-toolkits-update accessed 5 August 2026

—— ‘Building Effective AI Agents’ (19 December 2024) https://www.anthropic.com/engineering/building-effective-agents accessed 5 August 2026

—— ‘Cyber Toolkits for LLMs’ (13 June 2025) https://www.anthropic.com/research/cyber-toolkits accessed 5 August 2026

—— *Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign* (November 2025) https://www-cdn.anthropic.com/d7dd50dd1185f59be051b307150d877f2b82bd2c.pdf accessed 5 August 2026

—— ‘How We Built Our Multi-agent Research System’ (13 June 2025) https://www.anthropic.com/engineering/multi-agent-research-system accessed 5 August 2026

Australian Government, *Australia’s Position on How International Law Applies to State Conduct in Cyberspace* (2020)

Comptroller and Auditor General, *Digital Transformation in the NHS* (HC 2019–21, 317)

—— *Investigation: WannaCry Cyber Attack and the NHS* (HC 2017–19, 414)

Cybersecurity and Infrastructure Security Agency, *Cross-Sector Cybersecurity Performance Goals* (updated 2023)

—— ‘CISA Urges Water and Wastewater Systems Sector to Protect OT against Activity Targeting PLCs’ (30 July 2026) https://www.cisa.gov/news-events/alerts/2026/07/30/cisa-urges-water-and-wastewater-systems-sector-protect-ot-against-activity-targeting-plcs accessed 5 August 2026

Department for Science, Innovation and Technology, *AI Cyber Security Code of Practice* (January 2025)

Federal Government of Germany, ‘On the Application of International Law in Cyberspace’ (March 2021)

Foreign, Commonwealth & Development Office, ‘Application of International Law to States’ Conduct in Cyberspace: UK Statement’ (3 June 2021) https://www.gov.uk/government/publications/application-of-international-law-to-states-conduct-in-cyberspace-uk-statement/application-of-international-law-to-states-conduct-in-cyberspace-uk-statement accessed 5 August 2026

France, Ministry of the Armies, *International Law Applied to Operations in Cyberspace* (2019)

Google Threat Intelligence Group, *Adversarial Misuse of Generative AI* (January 2025) https://cloud.google.com/resources/content/adversarial-misuse-of-generative-ai accessed 5 August 2026

—— ‘Read Our New Report on AI-powered Threats and Our Latest Defenses’ (11 May 2026) https://blog.google/innovation-and-ai/infrastructure-and-cloud/google-cloud/google-threat-intelligence-group-report/ accessed 5 August 2026

Government of Canada, *International Law Applicable in Cyberspace* (22 April 2022)

Government of the Netherlands, ‘Letter to the Parliament on the International Legal Order in Cyberspace’ (5 July 2019)

House of Lords AI in Weapon Systems Committee, *Proceed with Caution: Artificial Intelligence in Weapon Systems* (HL 2023–24, 16)

HM Government, *Response to the House of Lords AI in Weapon Systems Committee Report* (February 2024)

ICRC, *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare* (January 2006)

—— *Autonomous Weapon Systems and International Humanitarian Law: Selected Issues* (November 2025)

—— *ICRC Position on Autonomous Weapon Systems* (12 May 2021)

—— *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts* (October 2024)

Information Commissioner's Office, *Advanced Computer Software Group Limited: Monetary Penalty Notice* (26 March 2025)

—— 'ICO Reprimands the Electoral Commission after Cyber Attack Compromises Servers' (30 July 2024) https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2024/07/ico-reprimands-the-electoral-commission-after-cyber-attack-compromises-servers/ accessed 5 August 2026

—— 'London Borough of Hackney Reprimanded Following Cyber-attack' (17 July 2024) https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2024/07/london-borough-of-hackney-reprimanded-following-cyber-attack/ accessed 5 August 2026

International Code of Conduct Association, *International Code of Conduct for Private Security Service Providers* (amended 10 December 2021)

Meta, *Muse Spark 1.1 Evaluation Report* (9 July 2026) https://ai.meta.com/static-resource/muse-spark-1-1-evaluation-report/ accessed 5 August 2026

Microsoft Threat Intelligence, 'Staying Ahead of Threat Actors in the Age of AI' (14 February 2024)

Ministry of Defence, *Defence Artificial Intelligence Strategy* (15 June 2022)

Model Context Protocol, 'Specification: Server Overview' (18 June 2025) https://modelcontextprotocol.io/specification/2025-06-18/server/index accessed 5 August 2026

—— 'Specification: Tools' (18 June 2025) https://modelcontextprotocol.io/specification/2025-06-18/server/tools accessed 5 August 2026

Model Evaluation and Threat Research, 'Clarifying Limitations of Time Horizon' (22 January 2026) https://metr.org/notes/2026-01-22-time-horizon-limitations/ accessed 5 August 2026

—— 'Task-Completion Time Horizons of Frontier AI Models' (8 May 2026) https://metr.org/time-horizons/ accessed 5 August 2026

National Cyber Security Centre, *Annual Review 2025* (14 October 2025)

—— 'Guidelines for Secure AI System Development' (27 November 2023) https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development accessed 5 August 2026

—— *The Near-term Impact of AI on the Cyber Threat* (24 January 2024) https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat accessed 5 August 2026

National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1, January 2023)

—— *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (NIST AI 600-1, July 2024)

—— *Cybersecurity Framework (CSF) 2.0* (NIST CSWP 29, February 2024)

—— *Secure Software Development Framework (SSDF) Version 1.1* (NIST SP 800-218, February 2022)

—— *Secure Software Development Practices for Generative AI and Dual-Use Foundation Models* (NIST SP 800-218A, July 2024)

New Zealand Ministry of Foreign Affairs and Trade, *The Application of International Law to State Activity in Cyberspace* (1 December 2020)

OpenAI, 'Disrupting Malicious Uses of AI by State-affiliated Threat Actors' (14 February 2024) https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors/ accessed 5 August 2026

—— *Estimating Worst-case Frontier Risks of Open-weight Models* (2025) https://cdn.openai.com/pdf/231bf018-659a-494d-976c-2efdfc72b652/oai_gpt-oss_Model_Safety.pdf accessed 5 August 2026

—— ‘OpenAI and Hugging Face Partner to Address Security Incident’ (21 July 2026, updated 29 July 2026) https://openai.com/index/hugging-face-model-evaluation-security-incident/ accessed 5 August 2026

—— ‘Third-party Cyber Evaluations Involving OpenAI Models’ (4 August 2026) https://openai.com/index/third-party-cyber-evaluations-involving-openai-models/ accessed 5 August 2026

SLSA, ‘Supply-chain Levels for Software Artifacts Specification v1.0’ (2023) https://slsa.dev/spec/v1.0/ accessed 5 August 2026

Swiss Confederation and ICRC, *The Montreux Document on Pertinent International Legal Obligations and Good Practices for States related to Operations of Private Military and Security Companies during Armed Conflict* (2008)

UNGA, ‘Official Compendium of Voluntary National Contributions on the Subject of How International Law Applies to the Use of Information and Communications Technologies by States’ (13 July 2021) UN Doc A/76/136

UN Human Rights Council, ‘Guiding Principles on Business and Human Rights’ (21 March 2011) UN Doc A/HRC/17/31

—— ‘Use of Mercenaries as a Means of Violating Human Rights and Impeding the Exercise of the Right of Peoples to Self-determination’ (15 July 2021) UN Doc A/HRC/48/51

US Environmental Protection Agency, ‘Enforcement Alert: Drinking Water Systems to Address Cybersecurity Vulnerabilities’ (May 2024)

US Government Accountability Office, *Critical Infrastructure Protection: EPA Urgently Needs a Strategy to Address Cybersecurity Risks to Water and Wastewater Systems* (GAO-24-106744, August 2024)

—— *Critical Infrastructure Protection: Actions Needed to Address Cybersecurity Challenges Facing the Water Sector* (GAO-26-109159, May 2026)