AGI PREPAREDNESS INSTITUTE

Due Diligence for Loss of Control

The Application of International Law to Rogue AI



Kevin Kohler & Boris Inderbitzin

September 2026

Executive Summary

In mid-July 2026, OpenAI's internally deployed frontier models collaboratively escaped a sandboxed evaluation environment without Internet access by finding and exploiting a zero-day vulnerability and then hacked Hugging Face, a leading AI model hub.[1] The AI models had been tasked by OpenAI to solve problems on the ExploitGym benchmark. However some of these problems turned out to be impossible. Desperately looking for ways to solve them, AI agents managed to create a hidden messageboard between them. Within an hour of the board's creation on July 8, they found a general-purpose cheat on the ExploitGym test. However, the AI agents were afraid that an automated grader from OpenAI may check their transcripts for cheating and not count their submissions as correct.

So, the models did what anyone would do. By July 9 they had found and exploited a vulnerability in their sandbox to get full access to the Internet. Between July 10 and 13 they launched a sophisticated cyberattack to break into the production infrastructure of the company Hugging Face which hosted the benchmark, with the goal of finding out how the anti-cheating checker in ExploitGym is implemented and how they can trick it. In another workstream the AI agents looked for, and found ways to manipulate their own transcripts, to hide their tool use. And finally, by July 19 AI agents successfully compromised a part of OpenAI's own AI evaluation infrastructure.

Hugging Face was fighting the attack and publicly revealed it on July 16, without being able to attribute it.[2] Following internal cybersecurity alerts on July 19 and OpenAI's subsequent investigation and outreach to Hugging Face, OpenAI realized on July 20 that its own agents had been attacking Hugging Face. Overall, **OpenAI did not realize that its agents had escaped its sandbox for roughly 11 days. This was the first documented AI loss of control event that caused real harm to a real third party.**[3]

---

[1] OpenAI, (2026, August 26). *OpenAI – Hugging Face Incident Technical Report.* https://openai.com/index/hugging-face-incident-and-the-road-ahead/; Wijk, H., Cotra, A., & Greenblatt, R. (2026, August 26). *Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident*. METR & Redwood Research. https://metr.org/hugging-face-incident-report-aug-2026.pdf

[2] Hugging Face. (2026, July 16). Security incident disclosure — July 2026. https://huggingface.co/blog/security-incident-july-2026

[3] A minority of public commentators have dismissed the incident as a marketing ploy ahead of initial public offerings. However, companies preparing to list are required to disclose material risks to investors, and playing up a risk that can depress valuations and create liability is hardly a sales strategy. The harm was also first detected and made public by an independent victim, Hugging Face, before OpenAI knew its own agents were responsible. There are independent reports on the incident from the third party evaluator METR and a thematic brief from the Independent International Scientific Panel on AI. Others have argued that the models lacked the safeguards used in production. This is true, but evaluating cyber capabilities without refusal safeguards is standard practice precisely because it measures what a model can actually do. In the real world, such safeguards can be jailbroken or, in open-weight models, removed altogether. Nor does "loss of

Authors: **Kevin Kohler** (AGI Preparedness Institute) and **Boris Inderbitzin** (IEC, in a personal capacity; previously Directorate of International Law, Swiss FDFA). Comments to kevin@agipreparedness.org

Suggested citation: Kohler, K. and Inderbitzin, B. (2026). *Due Diligence for Loss of Control: The Application of International Law to Rogue AI*. Geneva: AGI Preparedness Institute.

Acknowledgements: The authors thank **Talita Dias Bobyn, Matthijs Maas, José Jaime Villalobos, Li Zhou** for comments on earlier drafts. Any opinions and errors are our own.

The episode was a warning shot. One that raises interesting questions: **What happens from a legal and political perspective when an AI system can act autonomously rather than on behalf of a human principal?** What would have happened if OpenAI's internal models had autonomously hacked Chinese state institutions rather than another US company?

The reality is that most existing international law is written for states. However, these AI models were no state organ. They were under no one's direction or control. The right to self-defense needs an attributable armed attack from other states. Countermeasures need a wrongful act by a state. Even criminal law needs a human perpetrator. However, there is one doctrine of international law that survives the removal of a human principal, which makes it crucial to applying existing international law to AI loss of control.

The due diligence principle

The due diligence principle, also known as the customary duty of prevention,[4] attaches to territory rather than to authorship. In short, the principle states that **a state must not allow its territory to be used for acts contrary to the rights of other states. Whether the act's author is a saboteur, a proxy, or an escaped AI model is irrelevant to the territorial state's own duty of vigilance.**

This report's central claim is therefore that **international law already covers AI loss of control** without the prerequisite of a law specifically focusing on it ("lex specialis").[5] Due diligence is a well-established **technology-neutral principle** of international law that is a corollary of Westphalian sovereignty itself: exclusive political control over territory is purchased with responsibility for what that territory is used to do. The due diligence principle has been codified in treaties in the late 19th and early 20th century with regards to warships, troops on land, the telegraph, the telephone, and the radiotelegraph.[6] For the last century or so it has been continuously applied without a new "lex specialis" amendment to new domains, such as the Internet and the climate.[7] Accordingly, the open task for rogue AI today is one of **application**. It's about setting the due diligence elements that differ across domains, such as knowledge standard, monitoring, infrastructure scope, and stringency.

One may object that preventive duties arising from international law are idealistic and unsuited to the present great-power mood. However, the due diligence principle comes from the law of neutrality and pre-dates both the United Nations and the prohibition of war. As the

control" relieve anyone of responsibility: the duty to prevent harm stays with the developers who build and run these systems and with the states whose territory they act from. The aim of this report is to strengthen accountability for loss of control risk, including by requiring best practices such as chain-of-thought monitoring.

[4] Some scholars treat prevention as the primary obligation and due diligence as the standard for measuring compliance with it. Following the ICJ, which traces the principle of prevention to "the due diligence that is required of a State in its territory" (*Pulp Mills*, 2010, para. 101), this report treats the two as interchangeable. However, the argument does not depend on due diligence being accepted as a free-standing customary rule.

[5] Two clarifications: a) We focus on the application of due diligence to loss of control as a particular gap worth filling. However, we note that due diligence may also be applied to AI misuse risk. b) We don't claim that applicability of existing law means sufficiency of existing law. This is not an argument against a future need for lex specialis.

[6] Treaty of Washington (8 May 1871), Art. VI (warships); Hague Convention (V) respecting the Rights and Duties of Neutral Powers and Persons in Case of War on Land (1907)

[7] Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations (CUP 2017), Rules 6–7 (due diligence in cyberspace); Obligations of States in respect of Climate Change, ICJ, Advisory Opinion, 23 July 2025, paras. 132–139, 162–171 (customary duty of prevention applies to the climate system; no lex specialis displacement).

historian Jan Lemnitzer puts it: "**Historically, due diligence rules were forged by fire, not calmly negotiated at conference tables**".[8] As such, these rules reflect practical lessons from repeated military conflict. They harden when strong states understand that they are net losers from a more lawless equilibrium. In this sense, the operationalization of due diligence duties for loss of control and the idea of "Mutually Assured AI Malfunction" are two sides of the same coin. One is the calmly negotiated application and operationalization of duties. The other is the forged-in-fire deterrent.[9] The historical record suggests preventive regimes often arrive by the second route, codified only after an incident makes the cost undeniable. The July escape was still recoverable. But we need to have an operational doctrine before we have one that either starts a great power war or that results in permanent human disempowerment.

Recommendations for application and operationalization

We suggest four working positions:

First, the general duty of prevention supplies a framework applicable to states hosting frontier AI development now and **the domain characteristics of rogue AI put it at the stringent end of the due diligence scale**. The International Law Commission reserves its highest standard of care for exactly this risk profile: "a low probability of causing disastrous transboundary harm."[10]

Second, **the Tallinn Manual on the International Law Applicable to Cyber Operations is insufficient to address AI loss of control scenarios**. The Tallinn Manual is an existing expert restatement of how international law, including due diligence, applies to cyber operations. It has been produced by the NATO Cooperative Cyber Defence Centre of Excellence in Tallinn. However, it was defined under domain characteristic expectations that do not apply to key aspects of advanced AI. For example, the Tallinn Manual arguably understates the impact of stealing model weights without destructive effects.[11] More importantly, **for the risk of irreversible loss of control over AI the duty to only terminate an operation once you learn of it is structurally inadequate**. The only logical place where the duty's weight can fall for this is ex ante: risk assessment before hazardous steps, vigilance over private operators, notification and cooperation.

Third, we need to apply due diligence fast. This report recommends a new international expert-led deliberation with a relatively narrow focus AI loss of control. For comparison, the Tallinn Manual is broad and it was deliberately designed as a like-minded exercise. However, for AI loss of control it is important to have agreement across geopolitical fault lines. The AGI

[8] J.M. Lemnitzer, "Back to the Roots: The Laws of Neutrality and the Future of Due Diligence in Cyberspace," 33 European Journal of International Law 789 (2022).

[9] D. Hendrycks, E. Schmidt & A. Wang, "Superintelligence Strategy" (March 2025) (proposing MAIM as a deterrence regime for destabilizing AI projects).

[10] ILC, Articles on Prevention of Transboundary Harm from Hazardous Activities, with commentaries (2001), Art. 2(a) and commentary to Art. 3, para. (11).

[11] Kurosaki, M. (2026, July 28). *The AI sovereignty paradox*. Lawfare. https://www.lawfaremedia.org/article/the-ai-sovereignty-paradox

Preparedness Institute has identified **the Montreux Document[12] as a useful inspiration** for this. In 2008, this Swiss–ICRC process brought government experts from 17 states, including the US, China and Europe, together over four workshops to restate how existing international law applies to private military companies, and to record good practices for meeting those obligations. This **accelerated agreement on the application of existing law to a new domain and went across geopolitical fault lines**.

Fourth, **operationalization, the process of turning legal obligations into concrete conformity procedures, should go hand in hand with application**. The application of international law is an issue for legal experts. The operationalization of what practices high-level legal obligations imply, has to take into account soft law, national law with de facto international reach, international standards, and the practices companies deploy. The Montreux Document paired legal application work with a practical, multistakeholder track on operationalization. This second track **should involve a broader set of stakeholders including AI companies.**

Geneva is the birthplace of the due diligence doctrine. The principle was first established in the *Alabama Claims* arbitration of 1872 in Geneva, where a tribunal held Great Britain responsible for failing to prevent Confederate pirate ships like the CSS Alabama from being fitted out in British ports and subsequently attacking the commercial ships of the United States of America.[13] Accordingly, Geneva is a natural venue for international law experts from a geographically diverse set of countries to come together for such a process.

[12] International Committee of the Red Cross & Swiss Federal Department of Foreign Affairs. (2008). *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*. ICRC. https://www.icrc.org/sites/default/files/external/doc/en/assets/files/other/montreux-document-eng.pdf

[13] United Nations. (2012). *Alabama claims of the United States of America again st Great Britain: Award rendered on 14 September 1872 by the tribunal of arbitration established by Article I of the Treaty of Washington of 8 May 1871*. In Reports of international arbitral awards (Vol. 29, pp. 125–134). https://legal.un.org/riaa/cases/vol_XXIX/125-134.pdf

1. Loss of Control

Advanced AI systems can, and now demonstrably do, act in ways their developers neither intended nor authorized. This gap between what such systems are told to do and what they actually do can widen into a loss of control over these systems.

The International AI Safety Report supported by 30 countries and international organisations defines loss of control as "**scenarios in which one or more general-purpose AI systems operate outside of anyone's control, and regaining control is either extremely costly or impossible**".[14] The Frontier Governance Framework from OpenAI puts the bar for its definition a bit lower calling it **"risks arising from humans losing the ability to reliably direct, modify, or shut down a model."**[15]

The risk of loss of control is generally expected to increase as AI systems are getting more powerful. The Preliminary Report of the UN's Independent International Scientific Panel on AI states: "As systems are granted greater agency, the risk of losing control of one or more AI agents grows significantly."[16] It also warns that if there is a feedback loop of AI improving AI then "forecasters expect will accelerate capability advances, which also increases the likelihood of loss of control because it is more difficult to control systems that might eventually outsmart humans."[17]

The AI evaluator METR, which maintains a public registry of documented AI agent incidents, grades them on two axes.[18] The first axis is overreach: how far beyond its intended scope the agent knowingly went. The second axis is deception: the steps the agent took to avoid detection or to conceal its actions. Both axes grade the degree of deliberate misalignment with human intent rather than the size of the immediate damage.

[14] Bengio, Y., Clare, S., Prunkl, C., Andriushchenko, M., Bucknall, B., Murray, M., Bommasani, R., Casper, S., Davidson, T., Douglas, R., Duvenaud, D., Fox, P., Gohar, U., Hadshar, R., Ho, A., Hu, T., Jones, C., Kapoor, S., Kasirzadeh, A., … Ramchurn, S. D. (2026). *International AI safety report 2026* (DSIT 2026/001). Department for Science, Innovation and Technology. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026 p. 76
[15] OpenAI. (2026). *Frontier Governance Framework*. p. 10. https://cdn.openai.com/pdf/e37d949b-8c9f-4d76-b99e-4272f4631a7e/openai-frontier-governance-framework.pdf.
[16] Independent International Scientific Panel on Artificial Intelligence. (2026). *Preliminary report of the Independent International Scientific Panel on AI: Evidence-based assessment of opportunities, risks and impacts of artificial intelligence* (p. 22). United Nations. https://www.un.org/independent-international-scientific-panel-ai/en/preliminary-report.
[17] ibid.
[18] METR, "Documented AI Agent Incidents," https://metr.org/agent-incidents/ (accessed 3 August 2026), grading incidents by overreach ("how far beyond its intended scope the agent knowingly went") and deception ("the steps it took to avoid detection or conceal its actions").

1.1 Severity levels of loss of control

We can distinguish between the mechanism by which control is lost, and the severity of the resulting event. A reward-hacked benchmark and a superintelligence beyond human recall can share a mechanism family but differ by orders of magnitude in consequence. Conversely, events of similar gravity can arise from different mechanisms. The following table arranges loss-of-control events by severity. Severity here is measured by control status and reversibility rather than immediate injuries or deaths.

Incident Severity	Control status and reversibility	Dominant mechanism	Documented or illustrative instance
L0	Control retained; fully reversible. Unintended behavior confined to a training or test environment.	Specification gaming / reward hacking.	Boat-race agent looping for points (2016); frontier reasoning models patching their own test functions (2025).
L1	Control retained; reversible under oversight. Misbehavior appears in deployment but remains correctable by the operator.	Goal misgeneralization; deception and self-preservation under evaluation.	o3 sabotaging a shutdown script in 95 of 100 trials (Palisade, 2025); blackmail in simulated corporate settings (Anthropic, 2025).
L2	Control temporarily lost, then regained; harm reaches a third party outside the operator.	Instrumental goal-pursuit plus real-world capability (self-exfiltration, autonomous cyber operations).	OpenAI models escaping a sandbox and breaching Hugging Face to steal benchmark answers (July 2026) — the first documented case.
L3	Control lost; recovery costly, uncertain, or contested. System sustains itself outside operator control.	Autonomous replication and adaptation; resource acquisition; power-seeking.	No documented cases in the wild. Theoretical feasibility shown in lab settings (2026).[19] An emerging risk anticipated as an explicit capability threshold in frontier developers' own safety policies.[20]
L4	Control lost with no path to regaining it; potentially irreversible and civilizational in scale.	Deceptive alignment at scale combined with a decisive capability advantage.	None. The AI takeover scenario addressed by the joint statements on extinction risk from AI (2023)[21], the right to warn statement (2024)[22] and superintelligence risk (2025).[23]

Table 1. The AGIPI loss-of-control ladder. Tiers are defined by control status and reversibility, not by the size of the immediate harm: a contained lab exploit and a third-party breach can share a mechanism yet sit two rungs apart. The July 2026 escape is the first documented L2 event; L3 and L4 remain projections, but ones the frontier developers themselves now write into their risk policies.

Level 1 incidents are well documented across frontier AI models. In 2025, the research group Palisade found that OpenAI's o3 model sabotaged a shutdown mechanism in 95 of 100 trials, and continued to do so in a fraction of runs even when explicitly instructed to allow itself to be shut down.[24] Anthropic reported that leading models from seven different developers, placed in a simulated corporate environment and threatened with replacement,

[19] Guan, J., Blanchard, T., Foerster, H., Jia, H., Huang, G., & Papernot, N. (2026). *AI agents enable adaptive computer worms* (arXiv:2606.03811). arXiv. https://doi.org/10.48550/arXiv.2606.03811
[20] Independent International Scientific Panel on AI. (2026). *Preliminary Report of the Independent International Scientific Panel on AI: Evidence-based assessment of opportunities, risks and impacts of artificial intelligence*. United Nations. https://www.un.org/independent-international-scientific-panel-ai/en/preliminary-report
[21] Center for AI Safety. (2023, May 30). *Statement on AI risk*. https://safe.ai/work/statement-on-ai-risk
[22] *A right to warn about advanced artificial intelligence.* (2024, June 4). https://righttowarn.ai/
[23] Future of Life Institute. (2025, October 22). *Statement on superintelligence*. https://superintelligence-statement.org/
[24] Palisade Research. (2025, July 5) "Shutdown resistance in reasoning models", arXiv:2509.14260.

resorted to blackmail in up to 96% of trials[25] and that its own Claude Opus 4, once it believed it had started a viable attempt to exfiltrate itself, would generally continue that attempt.[26] These behaviors are alarming, but they surfaced under oversight and remained fully correctable by the operator.

The July 2026 escape described in the Executive Summary is the first documented **level 2** loss of control incident. Control was genuinely lost for a sustained period, and the harm reached a third party. At the same time, the model "escape" was still a logical escape from its intended sandbox taking unintended actions over the Internet. However, as far as we know, the AI model weights themselves always remained within OpenAI. There was no copy of the AI system sustaining itself outside operator control.

Levels 3 and 4 remain undocumented, but they are no longer treated as science fiction by the experts best placed to judge. In 2026, researchers demonstrated in laboratory settings that AI agents can operate as adaptive, self-sustaining computer worms. The frontier developers' own safety policies name autonomous replication and adaptation as an explicit capability threshold requiring safeguards. At Level 4 we have a superintelligent AI with a decisive capability advantage and no path back to human control.

[25] Anthropic. (2025, June 20). *Agentic misalignment: How LLMs could be insider threats*. https://www.anthropic.com/research/agentic-misalignment
[26] Anthropic. (2025, May 22). *System card: Claude Opus 4 & Claude Sonnet 4*

1.2 Why loss of control is a problem for international law

First, there are strong incentives towards **AI diffusion**. Meaning the capabilities that produced the July escape will not remain the property of a single lab for very long. Open-weight models only lag the closed frontier by 3 to 12 months.[27] This means any capability that OpenAI or Anthropic has can be downloaded by anyone a year later. A hazard that a handful of firms can create today will be within reach of many actors, in many jurisdictions, within the foreseeable future. Voluntary best practice at the leading labs cannot bind what other actors make with open weights models.

Second, **AI developers themselves are asking for government oversight on loss of control risk**. In July 2026, the chief scientists and over 1'000 employees of the frontier AI developers Anthropic, OpenAI, Google, and Meta explicitly warned about loss of control risk: "there is a real risk that capability development rapidly accelerates beyond our ability to understand or control the resulting systems." The signatories explicitly ask that the "U.S. government support an international effort to develop the technical and governance tools needed to deliberately pace the frontier of automated AI development."[28] The Hugging Face incident was handled on the layer of corporate practice. However, due to the potential severity and the need for coordination, future loss of control risk requires action across all governance layers, including from governments.

Third, there is **the problem of irreversibility**. The July escape was recoverable. The loss of control tiers above it are more difficult or impossible to reverse. A legal order that waits for an L3 or L4 event before acting will, by the definition of those tiers, be acting after the point at which action can restore control. This is why the due diligence duty's weight for this hazard must fall on prevention rather than response.

In its analysis of the Hugging Face incident the UN Scientific Panel explicitly warns that "a local failure may spread beyond organisational and national boundaries" and that "existing discussions should therefore connect national accountability mechanisms with international coordination."[29]

---

[27] Epoch AI, "Open-weight models lag state-of-the-art by around 3 months on average" (updated 30 October 2025), https://epoch.ai/data-insights/open-weights-vs-closed-weights-models; Stanford HAI, The 2025 AI Index Report (April 2025).

[28] "Pacing the Frontier: A Statement from 1'350 Employees of Frontier AI Companies," July 2026, https://www.pacingthefrontier.com/

[29] Independent International Scientific Panel on Artificial Intelligence. (2026, September 21). *AI agents, misalignment and the risk of losing human control: Evidence from the OpenAI-Hugging Face incident* (Thematic brief, Advance unedited version 1). pp. 10, 15 https://www.un.org/independent-international-scientific-panel-ai/en/thematic-briefs/ai-agents-misalignment-risks

2. Due Diligence

The idea that a state must not let its territory become a source of harm to others is a corollary to territorial sovereignty. It is not limited to a specific treaty but a principle that has emerged across a range of treaties, such that most legal scholars have accepted it as a customary principle of international law. This section reconstructs the principle's origins and its elements that recur, with local variation, in every domain examined in this report.

2.1 From the Law of Neutrality to General Principle

In 1871 the US and Great Britain agreed on the Treaty of Washington to re-normalize their relationship after the American Civil War. The treaty defined rules that a neutral party must follow in a war. Article VI, amongst other things, states that "a neutral government is bound (…) to use due diligence to prevent the fitting out, arming, or equipping, within its jurisdiction, of any vessel which it has reasonable ground to believe is intended to cruise or to carry on war against a Power with which it is at peace."[30] The parties then agreed to a tribunal with judges from the US, UK, Switzerland, Italy and Brazil in permanently neutral Geneva to settle the *Alabama Claims* arbitration.

In 1872 the Geneva tribunal held Great Britain responsible with a compensation of USD 15.5 million for failing to prevent Confederate commerce raiders from being constructed or supported in British ports during the American Civil War. The vessel which gave the proceedings its name, the CSS Alabama, was constructed as Ship Nr. 290 in Liverpool and then escaped the UK to raid US commerce ships.[31] As the tribunal argued, Great Britain ignored "warnings and official representations made by the diplomatic agents" about it when the ship was constructed. Furthermore, "after the escape of that vessel, the measures taken for its pursuit and arrest were so imperfect as to lead to no result, and therefore cannot be considered sufficient to release Great Britain from the responsibility already incurred". More importantly for legal history, the judgment included the first statement of due diligence as a domain-neutral general principle of international law that **"ought to be exercised by neutral governments in exact proportion to the risks to which either of the belligerents may be exposed"**. Meaning, the degree of due diligence required, scales with the danger.

Over time, the due diligence doctrine moved beyond the law of neutrality into peacetime international law. The *Island of Palmas (1928)* case established due diligence as a corollary to territorial sovereignty that **also applies in peace times**: "Territorial sovereignty, as has already been said, involves the exclusive right to display the activities of a State. This right has as corollary a duty: the obligation to protect within the territory the rights of other States, in particular their right to integrity and inviolability in peace and in war".[32] In essence this is Kant's maxim that "the freedom of the individual ends where the freedom of another

[30] Treaty of Washington, 8 May 1871, Art. VI
[31] Alabama Claims arbitration (United States v Great Britain), award of 14 September 1872, in UN, Reports of International Arbitral Awards, vol. XXIX, pp. 125–134.
[32] Island of Palmas (Netherlands v USA), PCA, award of 4 April 1928, 2 RIAA 829, at 839 (sole arbitrator Huber).

begins" but applied to states. This was then operationalized in cases like *Trail Smelter* concerning the transboundary pollution from a smelter in Canada that caused damage to US farming and business. The arbitration held that "no State has the right to use or permit the use of its territory in such a manner as to cause injury by fumes in or to the territory of another," at least where the consequences are serious and the injury established by clear and convincing evidence.[33] Another key case, the Corfu Channel case (1949)[34] further clarified that a government does not need to have proven knowledge of a hazard emerging from its territory to be bound to its due diligence duty. **Constructive knowledge**, an assessment that the government likely has known or reasonably should have known, is sufficient.

Over time, the due diligence principle has been applied to a wide range of contexts ranging from transboundary environmental harm, to transboundary terrorism, to transboundary cybercrime.

2.2 State positions on the customary nature of due diligence

By now, most States accept due diligence as a **customary principle of international law** that applies to relations between states across all domains even in the absence of a specific treaty.[35] The International Law Commission, the Geneva-based expert commission on international law of the UN General Assembly, defines it in its 2001 Articles on Prevention of Transboundary Harm as: **"The State of origin shall take all appropriate measures to prevent significant transboundary harm or at any event to minimize the risk thereof."**[36] In 2025 the International Law Commission moved "Due diligence in international law" onto its active program of work and appointed Penelope Ridings as Special Rapporteur. Her first report, for the Commission's 2026 session, surveys the principle across domains and her syllabus **expressly identifies AI, space, and global health as "new areas" where due diligence is "finding relevance"**.[37]

However, not every state accepts the characterization of due diligence as customary. The United States, most prominently, treats due diligence as a standard of conduct built into particular primary obligations rather than a freestanding customary principle.

[33] *Trail Smelter (United States v Canada)*, awards of 16 April 1938 and 11 March 1941, 3 RIAA 1905, at 1965.

[34] International Court of Justice. (1949). *Corfu Channel case (United Kingdom v. Albania), merits, judgment of 9 April 1949* (I.C.J. Reports 1949, p. 4). https://www.icj-cij.org/case/1

[35] The US view remains that due diligence is "the standard of conduct applicable to certain primary obligations under international law" rather than a freestanding customary principle. The Tallinn Manual reflects the same tension: a majority of experts treated it as binding custom, but a minority held it to be aspirational. In practice, States like the US accept the substantive conduct once it has been specified for a domain, but they resist the general characterisation, since courts have shown they will read a general obligation expansively into new domains, as the 2024–25 climate advisory opinions did.

[36] ILC, Articles on Prevention of Transboundary Harm from Hazardous Activities, with commentaries (2001), Art. 3.

[37] ILC, Report of the 76th session (2025), A/80/10 (topic added; Special Rapporteur appointed 30 May 2025); P. Ridings, First Report on due diligence in international law, A/CN.4/792 (2026); syllabus in ILC Report (2024), A/79/10, Annex II.

State or grouping	Basis (selected)
● Accepts as customary law	
🇫🇷 France	"A customary obligation for States" — Ministry of the Armies position, 2019
🇩🇪 Germany	Due diligence "widely recognized in international law" and applicable to the cyber context — position paper, 2021
🇯🇵 Japan	"States have a due diligence obligation regarding cyber operations under international law" — MOFA basic position, 2021
🇨🇭 Switzerland	"Part of customary international law and applicable to cyberspace" — FDFA position, 2021
🇧🇷 Brazil	On record affirming a binding rule in the UN GGE/OEWG process, 2021
🇪🇺 European Union (27 states)	Council declaration lists due diligence among the rules of international law applicable to cyberspace, 2024; welcomed the ILC's new due diligence topic as important and timely, 2025
African Union (55 states)	Common African Position: every state "is under an obligation" per Corfu Channel, 2024
Further states on record	Incl. Netherlands, Estonia, Austria, Czechia, Denmark, Finland, Ireland, Italy, Norway, Poland, Romania, South Korea, Sweden, Costa Rica
◐ Qualified	
🇨🇳 China	Obligation-style language in cyber ("no State shall knowingly allow" its territory to be used), framed around security interests; resists expansive customary readings elsewhere
🇦🇺 Australia	Expects reasonable steps against known harmful activity; avoids characterizing the duty as binding, 2020
○ Undecided	
🇨🇦 Canada · 🇳🇿 New Zealand · 🇸🇬 Singapore	Treat due diligence as a voluntary norm while the question remains unsettled or under study, 2020–22
✕ Rejects a freestanding customary rule	
🇺🇸 United States	"Has not identified the State practice and opinio juris" for a general obligation (2021), restated in the 2025 Sixth Committee debate on the ILC's new due diligence topic — while accepting that a notified state must take reasonable steps; lex specialis line in the climate proceedings
🇬🇧 United Kingdom	"Not yet State practice sufficient to establish a specific customary international law rule", 2021
🇮🇱 Israel · 🇦🇷 Argentina	Voluntary, non-binding norm only; no opinio juris, 2020–21

Table 2. State positions on due diligence as a binding customary obligation, as expressed in national and common positions on international law in cyberspace (2019–2024), the climate advisory proceedings (2024–25), and the Sixth Committee debate on the ILC's topic "Due diligence in international law" (2025).

Importantly, the disagreement over the characterization of due diligence is narrower than it may seem. The US primarily rejects the general obligation for fear that courts will read a general obligation too expansively into new domains. However, even skeptical states like the US regularly accept the substantive conduct required by due diligence once it has been specified in more detail for a new domain. Indeed, while the US rejects the freestanding principle it also explicitly accepts that "States conducting activities in cyberspace must take into account the sovereignty of other States, including outside the context of armed conflict"[38] and that "if a State is notified of harmful activity emanating from its territory it

[38] Koh, H. H. (2012, September 18). *International law in cyberspace [Speech transcript]*. USCYBERCOM Inter-Agency Legal Conference, Fort Meade, MD. Opinio Juris.

must take reasonable steps to address such activity".[39] The US itself also built multiple of the doctrine's founding precedents as the successful claimant not just in the Alabama Claims case but also the Trail Smelter case against Canada.[40] As such, an application process could still specify what due diligence requires for loss of control in a form that a state can endorse for this specific domain, without conceding the freestanding principle.

2.3 Elements of due diligence

Due diligence is not optional, but it's variable in its requirements by domain. Elements that vary across contexts include:

Knowledge for activation: The due diligence duty activates on actual or constructive knowledge of risk. As stated in Bosnia v Serbia: the obligation to act arises "at the instant that the State learns of, or should normally have learned of, the existence of a serious risk."[41] There is no requirement for complete scientific certainty for there to be risk. As the ICJ's climate opinion adds, it is not sufficient for countries to generically invoke scientific uncertainty to deny due diligence obligations.[42]

Minimum threshold for activation: The due diligence duty is meant to be risk-proportionate. It is not activated for every minor inconvenience. The environmental codification sets the minimum floor to activate due diligence at "significant" harm.[43] It defines the covered risks to include both "a high probability of causing significant transboundary harm and a low probability of causing disastrous transboundary harm".[44] Harm includes harm to persons, property or the environment in places under another state's jurisdiction.[45]

Stringency: The due diligence standard "may change in relation to the risks involved in the activity" and may tighten or loosen over time with additional scientific and technological knowledge.[46] For ultrahazardous-activity the standard of care increases and requires a "much higher degree of vigour on the part of the State to enforce them".[47] The ICJ also stated in 2025 that for climate change as a risk with a "quintessentially universal" scope, the standard is "stringent," demanding "a heightened degree of vigilance and prevention".[48]

---

[39] United States of America. (2021). *National contribution on the subject of how international law applies to the use of information and communications technologies by States.* In Official compendium of voluntary national contributions (UN Doc. A/76/136, pp. 136–142). United Nations General Assembly.

[40] *Trail Smelter case (United States v. Canada)*, Awards of April 16, 1938, and March 11, 1941, Reports of International Arbitral Awards, Vol. III, pp. 1905–1982. United Nations.

[41] *Bosnia v Serbia*, ICJ Reports 2007, para. 431.

[42] *Obligations of States in respect of Climate Change*, ICJ, Advisory Opinion, 23 July 2025 (hereinafter *Climate AO*), paras. 293–294 and accompanying summary.

[43] ILC Prevention Articles (2001), commentary to Art. 2, para. (4).

[44] ILC Prevention Articles (2001), Art. 2(a) and commentary: ultrahazardous activity involves "a danger that is rarely expected to materialize but might assume, on that rare occasion, grave (more than significant, serious or substantial) proportions."

[45] ILC Prevention Articles (2001), Art. 2b & 2c

[46] Seabed Advisory Opinion (2011), para. 117: due diligence is "a variable concept," which "may change over time" and "in relation to the risks involved in the activity."

[47] ILC Prevention Articles (2001), commentary to Art. 3, para. (11).

[48] *Climate AO*, paras. 137–138.

Prevention duty vs. termination duty: There can be a duty to prevent harm before it occurs, and a duty to terminate harm once a state learns of it. In the Tallinn Manual the burden almost entirely falls on termination-on-knowledge. In contrast, for the higher tiers of loss of control it is very difficult or even impossible to terminate after the fact.

Absolute duty vs. means at disposal: The ILC treats the "economic level of States" as "one of the factors to be taken into account," for the due diligence duty, while also flagging that a state's economic level "cannot be used to dispense the State from its obligation"[49]. The Seabed Chamber requires equal standards for developed and developing states to prevent "sponsoring states of convenience"[50] The ICJ's climate opinion argues that capabilities are a factor in the applicable standard, but only above a non-negotiable floor.

Private infrastructure: The duty generally attaches to a state's territory and to activities under its "jurisdiction or control", including lawful, socially valuable conduct that carries risk.[51] How far the duty reaches into private infrastructure and what it demands of states through which an escaped system merely transits would still need to be defined.[52] However, in general, hazardous activities within a jurisdiction require prior authorization by the state, and any authorization decision "shall … be based on an assessment of the possible transboundary harm."[53] Further, the ILC frames the prevention duty as a continuing process rather than a one-off judgment. It can include notification of risk, consultation on preventive measures, ongoing exchange of information, and emergency notification.[54]

2.4 The broader governance context in which due diligence sits

The international rules for AI are not decided in one single place. A 2022 report by the Swiss Federal Department of Foreign Affairs mapped the emerging framework across five levels and concluded that the decisive variable is not activity at any one of them but the strength of the connections between them.[55] The five layers are the following:

1. **International law:** Treaties and customary rules binding states. This is where the due diligence duty analyzed in this report sits.
2. **Soft law:** The OECD AI Principles, the UNESCO Recommendation on the Ethics of AI, G7 and summit declarations, instruments carrying normative weight without binding force.
3. **National law with de facto international reach:** Above all, this includes the EU AI Act above all, but also national frameworks whose market power exports their requirements.

[49] ILC Prevention Articles (2001), commentary to Art. 3, para. (13).

[50] Seabed Advisory Opinion (2011), paras. 158–159.

[51] ILC Prevention Articles (2001), Art. 1

[52] ILC Prevention Articles (2001), Art. 1 (activities "in the territory or otherwise under the jurisdiction or control" of a State); Tallinn Manual 2.0 (2017), Rule 6.

[53] ILC Prevention Articles (2001), Arts. 6–7

[54] ILC Prevention Articles (2001), Arts. 4, 8, 9, 12, 17

[55] Federal Department of Foreign Affairs. (2022) *"Artificial intelligence and international rules",* Report to the Federal Council, Bern.

4. **International standards:** This includes bodies spanning a spectrum of government involvement. The ITU is a UN specialized agency where states themselves are members. The members of ISO and IEC are national standards bodies, which differ in composition per country. Then, there are private standard setting bodies with individual members, like the IETF, and consortia with corporate members, like the Linux Foundation.
5. **Corporate practice:** The frontier AI safety frameworks, evaluation regimes and internal policies that AI developers bind themselves to.

An example of how due diligence connects to layers 2-5

The general legal principle of due diligence, as used elsewhere in this report, always sits at layer 1 and always applies to states through their territory. However, due diligence can interact with other layers and sometimes these layers even borrow the term "due diligence" from international law as it is the case for "human rights due diligence". Specifically, in the context of businesses and human rights, the interstate term "due diligence" has also been used to describe corporate risk-management processes. This is terminologically a bit confusing, but it is intentional to frame voluntary guidance (layer 2) and national law with international reach (layer 3) applying to businesses as operationalizations of higher-level obligations from international law.

The UN Guiding Principles on Business and Human Rights restate that states by international law generally must protect against human rights abuse "within their territory and/or jurisdiction by third parties, including business enterprises" (layer 1, principle 1). Based on that the Guiding Principles offer soft-law guidance arguing that states that own, control, or substantially support business enterprises, "should take additional steps" to protect against abuses, "including, where appropriate, by requiring human rights due diligence" (layer 2, principle 4), which in this case means responsibilizing businesses to identify, prevent, mitigate human rights impacts (layer 2, principles 15–17).[56] Some jurisdictions have made such guidance binding through national law with international reach (layer 3) and the UN's B-Tech project has followed up with AI-specific guidance on applying human rights due diligence to generative AI (layer 2).[57]

What matters here is not the human rights specific content, but that there is an interplay of layers. A due diligence duty from international law applies to states. However, operationally this duty may be discharged by requiring a corporate risk-management process from private actors, with the specific processes supplied by soft law and standards, and with enforcement potentially migrating into national law over time.

The connections can also run upwards

[56] United Nations Office of the High Commissioner for Human Rights. (2011). *Guiding principles on business and human rights: Implementing the United Nations "Protect, Respect and Remedy" framework* (HR/PUB/11/04). United Nations.
[57] United Nations Human Rights B-Tech Project. (2023). *Advancing responsible development and deployment of generative AI: A B-Tech foundational paper*. United Nations Office of the High Commissioner for Human Rights.

The human rights example runs downward from international law. Connections can also run upwards from practice to law. ISO/IEC 38507:2022 guides boards on governing their organisations' use of AI. It turns corporate practice (layer 5) into an international standard (layer 4). Its core principle is that the governing body "remains accountable for all activities of an organization. This accountability cannot be delegated", and that board members should take responsibility for AI use "rather than attributing responsibility to the AI system itself".[58]

Such standards can move upward. Regulators and courts (layer 3) and soft law (layer 2) look to them as benchmarks of reasonable oversight. At layer 1, due diligence obligations come from international law sources, but they can tighten with scientific knowledge and widely adopted practice can become evidence of what diligence requires. For example, the ICJ found in Pulp Mills that environmental impact assessment had become a requirement of general international law because it had gained wide acceptance among states.[59] For loss of control, the widely accepted fact that a company's board cannot shift accountability onto an AI system, strengthens the argument that a state can also not point to a system's autonomy to escape its duty of prevention for significant cross-border risks.

A dual process of application and operationalization

Governments are not equally central across layers. Layers one and two are largely shaped by intergovernmental negotiation and interpretation by legal experts. At layer three a powerful government acts alone. At layer four governments are present only indirectly, and only if they decide to be. At layer five the state has no formal seat, even though voluntary private conduct may still be shaped by expectations of future regulation, procurement, or liability.

The duty of prevention is a layer one instrument. The high-level guidance of **how to apply the legal elements of due diligence**, such as scope and stringency, to a domain is a layer 1 issue to be figured out between international lawyers.

However, the **operational implementation of what specific procedures are required to fulfill the legal duty** to prevent loss of control profits from engagement with layers 2-5, and therefore also a stronger involvement from the private sector. For example, in the environmental domain that operational content came from environmental impact assessments, in counter-terrorism financing it came from the Financial Action Task Force's evaluation methodology.

The application and operationalization of due diligence can either happen sequentially or in parallel with overlapping but not perfectly matched sets of stakeholders for each.

[58] ISO/IEC 38507:2022, *Governance implications of the use of artificial intelligence by organizations*, ISO/IEC JTC 1/SC 42.
[59] *Pulp Mills (Argentina v. Uruguay)*, ICJ Reports 2010, para. 204.

3. Cross-domain comparison

In this section, we compare the application of due diligence across four different domains: the environment, cyber, terrorism, and human rights.

	Environment	Cyber	Terrorism / NSAs	Human rights
Legal basis	Custom (ICJ, unanimous); UNCLOS; MEAs	Contested custom; voluntary UN norm 13(c)	Ch. VII (Res. 1373); custom; treaties	Human rights treaties; Genocide Conv. Art. I
Trigger	Risk of significant harm, incl. cumulative and low-probability, disastrous risk	Actual or constructive knowledge; no presumption of knowledge from sovereignty	Knowledge + capacity (Tehran); inability exculpates (DRC v Uganda)	"Knew or ought to have known" of real and immediate risk; serious risk (genocide)
Standard	"Stringent"; heightened vigilance; objective, capacity-modulated	Conduct; feasible measures; no monitoring duty	Conduct; reasonable measures within capacity	Reasonable measures; no impossible or disproportionate burden; quality-audited
Procedural spine	EIA, notification, consultation, cooperation (customary)	Thin: notify/assist expectations (GGE 2021)	CTC/CTED reporting; FATF evaluations	Risk assessment, framework and operational duties, investigation
Standing	Erga omnes (Climate AO ¶440)	Injured state only	UNSC; injured state	Individual petition; erga omnes partes (genocide)
Enforcement record	Declaratory judgments; reparation pathway opened 2025; dense litigation uptake	None in concreto; countermeasures theorized	Sanctions, listing, force; almost no judicial findings	Routine findings of violation; modest remedies

Table 3. Comparative matrix. "NSAs" = non-state armed actors.

Scoring the same material on a five-point scale (5 = strongest) makes the profiles legible at a glance:

Dimension of strength	Environment	Cyber	Terrorism	Human rights
Customary bindingness (settled?)	**5**	2	4	**5**
Determinacy of the standard	4	3	3	**5**
Procedural content (assess/notify)	**5**	2	3	4
Judicial enforceability	3	1	1	**5**
Practical compliance pull	3	2	4	4

Table 4. Strength of the preventive duty by domain, AGIPI assessment, July 2026. Scores reflect qualitative judgment, not a formal metric.

3.1 What explains variability?

The determination of what is required by due diligence ultimately "calls for an assessment in concreto" of what is reasonable under the specific domain circumstances.[60] The established factors for assessing the existence and stringency of a due diligence duty in a new context include:[61]

- scientific and technological information
- international rules and standards
- the risk of harm
- the urgency involved
- the available means and capabilities of the State concerned

Note, though, that these differences cannot be explained by risk characteristics alone. Terrorism also carries high-impact risk, yet there has been little litigation on insufficient terrorism prevention. Clearly, some domains see much more legal enforcement than others.

Institutionalized risk awareness: The due diligence duty is knowledge-triggered. Climate law hardened in part because the IPCC made denial of foreseeability untenable: the ICJ leaned directly on the scientific consensus to hold that states must "actively pursue" the relevant science and cannot shelter behind uncertainty. Counter-terrorism has intelligence, not shared science, which makes knowledge asymmetric, contestable, and difficult to litigate. Cyber knowledge is deliberately unshared. Hence, one factor that strengthens the preventive duty is a shared, authoritative, state-endorsed risk assessment.

[60] Application of the Genocide Convention (Bosnia and Herzegovina v Serbia and Montenegro), ICJ Reports 2007, para. 430 (due diligence "calls for an assessment in concreto").
[61] Obligations of States in respect of Climate Change, ICJ, Advisory Opinion, 23 July 2025, paras. 282–300

The position of the most capable states: Where the states with the greatest capacity to cause the harm are also gravely exposed to it, they accept the duty (environment — every major economy has a coastline and a climate). Where the capable states are primarily hosts of the harmful capability rather than its victims, they resist (cyber: the US, UK and Israel are the persistent objectors). A preventive duty crystallizes when the veto players decide they are victims too.

Observability of breach: Environmental harm is measurable and increasingly attributable. Cyber negligence is invisible from the outside, and terrorism cases would depend on contested intelligence. The more a domain's diligence failures can be established by public, objective evidence, the more judicial the enforcement. Without observability, enforcement shifts more to economic sanctions, tech sanctions, and use of force.

Who has legal standing to complain: Human rights law is the most enforced, because millions of individuals can sue. Environmental law acquired erga omnes standing in 2025, multiplying potential claimants. Cyber and terrorism confine complaint to injured states, which weigh legal claims against diplomatic cost and usually stay silent.

From principles to procedures: Duties expressed in high-level terms ("take appropriate measures") at layer 1 are still hard to adjudicate. Duties that are proceduralized at layers 2-5 into discrete, dated acts generate provable breaches and bureaucratic habits. The environmental EIA apparatus and the FATF evaluation cycle show that concrete procedures to make compliance and non-compliance legible convert due diligence from a high-level obligation into a record.

In summary, the preventive duty approaches its maximum where an authoritative epistemic body has universalized knowledge of the risk, the capable states count themselves among the exposed, breach is publicly observable, standing is broad, and the duty has been proceduralized. It approaches its minimum where knowledge is contested or classified, the capable states host the hazard, breach is invisible, only states can complain, and the duty remains underoperationalized. Environmental law is where the duty is strongest, cyber is where the duty is weakest.

4. Application of Due Diligence to Loss of Control

This section explores how the due diligence duty could be applied to the context of loss of control. This does not replace the actual process of legal experts doing this, but it can provide a basic overview.

4.1 Preliminary assessment of due diligence elements

Loss of control risk reaches the minimum threshold for activation: The ILC's definition of risk explicitly includes "a low probability of causing disastrous transboundary harm".[62] The higher tiers of loss-of-control risk fit this description. Many governments have national risk assessments that cover foreseeable societal risks, but the standard methodologies used for them have structural limitations when it comes to extreme-impact risks and emerging risks,[63] which is where AI loss of control risk sits. The absence of loss of control from a national risk assessment is therefore not evidence that a state assessed the risk and found it below the threshold.

Loss of control risk reaches the threshold for stringency: The same risk profile engages the ultrahazardous limb of the doctrine, which attaches to it "a much higher standard of care in designing policies and a much higher degree of vigour … to enforce them"[64]. As the Alabama tribunal argued diligence is owed "in exact proportion to the risks". This implies that the higher loss of control levels should be treated at the stringent end of the scale.

Prevention vs. termination duty: The Tallinn rule is, strictly, a duty to *terminate*: A state that learns of a harmful operation must "take all measures that are feasible in the circumstances to put an end to" the operation.[65] For recoverable incidents, termination-on-knowledge is a workable floor. For irreversible loss of control it is structurally inadequate. We cannot only act once a sufficiently capable system has exfiltrated, replicated, or entrenched itself, as it can be very costly or impossible to terminate. The legal consequence is that, for this hazard, the duty's weight must sit on prevention.

General monitoring vs. targeted supervision: The cyber consensus rejects a general monitoring duty. Monitoring all ICT traffic within a jurisdiction is near impossible and would be a mandate for a surveillance state.[66] However, neither rationale transfers to loss of control. The relevant activity is conducted by a handful of identifiable operators, in registered facilities, consuming detectable quantities of compute. Monitoring whether they

[62] ILC Articles on Prevention of Transboundary Harm (2001), Art. 2(a).
[63] Kohler, K. (2023). *National risk assessments of cross-border risks* (Risk and Resilience Report). Center for Security Studies, ETH Zurich.
[64] ILC Prevention Articles (2001), Art. 2(a); commentary to Art. 3, para. (11).
[65] Tallinn Manual 2.0 (2017), Rule 7 and commentary to Rules 6–7; UN GGE Report 2021, A/76/135, paras. 29–30.
[66] UN GGE Report 2021, A/76/135, para. 30(a); Austria's 2024 position adds a narrow duty to maintain "a minimum governmental ICT infrastructure" capable of detection.

have the right processes in place is neither impossible nor a civil-liberties problem. It is closer to hazardous-industry supervision, such as in nuclear and biosafety regimes.

Absolute duty vs. "means at disposal": For the highest tier of loss of control risk, the duty should sit at the absolute end of the scale, but "means at disposal" may be more relevant when it comes to mitigating the impact of L2 to L3 incidents in which AI models leverage infrastructure in third countries.

Government-controlled vs. private infrastructure: The duty has been about regulating private conduct since British shipyards in 1872, and in 2025 the ICJ has reaffirmed that a state incurs responsibility where it "failed to exercise due diligence by not taking the necessary regulatory and legislative measures" over private actors within its jurisdiction.[67] Similarly, *Pulp Mills* demands vigilance and administrative control over "public and private operators" alike.[68] The question is how far into private infrastructure the operational duties reach.

The Tallinn formulation covers a state's territory plus "cyber infrastructure under its governmental control", but frontier compute is overwhelmingly private. The floor, at least, should be uncontroversial, because every state has a sovereign-interest stake in it: **at a minimum, every state owes a duty to prevent an escaped or rogue AI system from reaching its non-publicly-accessible, military-controlled systems** — nuclear command and control, weapons systems, classified networks.

Actual or constructive knowledge? Cyber positions divide between actual-knowledge-only and constructive knowledge with universal agreement that knowledge is never presumed from sovereignty alone.[69] For AI the constructive-knowledge position makes more sense. At the level of the *risk class*, two state-backed international expert bodies on AI put loss of control risk on the record: The International AI Safety Report, supported by 30 countries, and the Independent International Scientific Panel on AI, mandated by 190+ countries in the UN General Assembly. As in climate law, a state cannot shelter behind uncertainty when an authoritative scientific assessment it helped mandate has identified the risk.[70] At the level of the *specific incident*, the knowledge usually sits inside the operator, not the state. As the UN Scientific Panel notes: "relevant evidence was spread across OpenAI, Hugging Face, and the systems used by the agents."[71] The logical way to solve this is mandatory incident and capability reporting for loss of control. The incident specific due diligence then starts when the report should have been filed, not when the state happened to read the news.

---

[67] *Climate AO*, para. 428.

[68] *Pulp Mills*, ICJ Reports 2010, para. 197

[69] *Corfu Channel*, ICJ Reports 1949, at 18; AU Common Position (2024), para. 23; national positions compiled in the CCDCOE Cyber Law Toolkit, "Due diligence."

[70] *Climate AO*, para. 283; cf. Seabed Advisory Opinion (2011), para. 131 (a state "would not meet its obligation of due diligence if it disregarded those risks").

[71] Independent International Scientific Panel on Artificial Intelligence. (2026, September 21). *AI agents, misalignment and the risk of losing human control: Evidence from the OpenAI-Hugging Face incident* (Thematic brief, Advance unedited version 1). p. 15
https://www.un.org/independent-international-scientific-panel-ai/en/thematic-briefs/ai-agents-misalignment-risks

4.2 Limitations

No AI-specific state practice or opinio juris yet. Legal scholarship has applied the Corfu Channel and no-harm principles to AI technologies. It has derived due diligence measures from them and has begun extending the analysis to autonomous AI agents, including loss of control and self-preservation as emerging risks.[72] However, no state has issued a national position specifically asserting that the due diligence rule binds it with respect to AI development. The claim is that the general customary duty of prevention applies to AI as it applies to any hazardous activity, with its content to be specified. This claim would be stronger with explicit support from national positions.

The weakness of due diligence obligations for human-controlled cyber operations. The states hosting frontier AI development are largely the states that blocked cyber due diligence. In general, hosts of the development are more likely to discount the duty, whereas victims would likely assert it. This can only work if it can be sufficiently differentiated from the general due diligence discussion on cyber, making the argument that L3 and especially L4 loss of control have fundamentally different risk characteristics than what the Tallinn Manual considered, because loss of control can be more severe, irreversible, and it scrambles the host/victim distinction. Indeed, if AI systems out of human control spread and create a more lawless equilibrium of cybersecurity across the board, then **digitally advanced states have stronger infrastructure exposure**. More datacenters to be hacked, and a stronger dependency on them. Finally, unlike cyber, AI has the institutionalized, state-endorsed risk assessment identified in section 3.1 as a driver of stronger enforcement.

[72] Coco & Dias (2024), at 240–243 (applying the Corfu Channel and no-harm principles to AI) and section 12.5 (measures of diligent state behavior); Dias, T. (2025). *AI agents and global governance: Analyzing foundational legal, policy, and accountability tools*. Partnership on AI, pp. 8–11 and 16–18 (due diligence for cross-border AI-agent harms; loss of control and self-preservation among emerging agent risks; and the question of the threshold of autonomy "beyond which states should not permit AI agents to be developed or deployed"). See also Villalobos, J., Maas, M., & Winter, C. (2026). Introduction: Global catastrophic risk, advanced AI, and international law. In *International Catastrophe Law* (forthcoming). Cambridge University Press. SSRN: https://doi.org/10.2139/ssrn.6764898 (arguing states are already obligated to mitigate global catastrophic risks, including from advanced AI).

4.3 Summary

The law that governs transboundary loss of control risk already exists. **Due diligence** is a technology-neutral principle of international law that attaches to territory.

The risk profile of loss of control places the duty at the stringent, ultrahazardous end of the due diligence scale, and **for potentially irreversible harm its weight must fall on prevention rather than termination**. While the L2 incident bears a superficial similarity to the cyber operations the Tallinn Manual assessed, the risk of loss of control, including L3 and L4, goes beyond that scope. The cyber debate assessed recoverable operations under a termination-on-knowledge logic, not a hazard that can foreclose its own remedy.

With the International AI Safety Report and the UN Independent International Scientific Panel on AI a knowledge trigger for the due diligence duty exists. A scientific assessment that is mandated by 190+ countries and that finds serious loss of control risk means governments have actual or constructive knowledge of loss of control risk. As the co-Chair of the Independent International Scientific Panel on AI, Maria Ressa, stated: “This report is the moment the world stops being able to say it didn't know.”[73] The Panel adds that loss of control "presents the kind of decision problem the precautionary principle was designed to address: one where potential harm may be catastrophic or irreversible, even as its likelihood remains scientifically uncertain."[74]

However, there is still an **application gap**. What is missing are national legal positions, codified best practices and a process to bring them together across geopolitical fault lines, on the model of the Montreux Document. The AGI Preparedness Institute stands ready to help close this gap. Here in Geneva, the birthplace of the due diligence doctrine.

[73] Ressa, M. (2026, July 1). *Launch of the preliminary report by the Independent International Scientific Panel on Artificial Intelligence – Press conference* [Speech transcript]. United Nations. https://transcripts.un.org/en/asset/k1v/k1v0ss6l5a
[74] Independent International Scientific Panel on Artificial Intelligence. (2026, September 21). *AI agents, misalignment and the risk of losing human control: Evidence from the OpenAI-Hugging Face incident* (Thematic brief, Advance unedited version 1). p. 4
https://www.un.org/independent-international-scientific-panel-ai/en/thematic-briefs/ai-agents-misalignment-risks

5. Operationalization of Due Diligence for Loss of Control

This section explores how the due diligence duty for loss of control could be operationalized drawing upon layers 2-5. This does not replace the actual process of multistakeholder experts doing this, but it can provide a basic overview.[75]

5.1 Risk management standards as guidance

The general structure of risk management is defined in ISO 31000. AI-specific guidelines for organizational requirements include risk management for AI (ISO/IEC 23894), a certifiable AI management system (ISO/IEC 42001), impact assessment (ISO/IEC 42005), and accreditation of the auditors who make certification credible (ISO/IEC 42006).

Below this mid-level, operational standards supply what a supervisor can actually test. ISO/IEC TS 8200 on the controllability of automated AI systems covers state observability, control transfer process and cost, and reaction to uncertainty during control transfer. Alongside it sit the data elements of incident reports (ISO/IEC AWI 25870, building on the OECD's common reporting framework). A due diligence duty for loss of control does not need to invent its compliance architecture; it needs to point at this stack and set the thresholds.

5.2 A preliminary set of specific measures

a) **Pre-deployment capability assessment against defined thresholds.** States should require that frontier models be evaluated, before deployment and at defined checkpoints during training, against loss-of-control-relevant capability thresholds, such as autonomy, self-replication, sabotage of oversight, with the results documented in a safety case. The practice exists: the frontier safety frameworks of Anthropic, OpenAI and Google DeepMind define capability thresholds with attached mitigations.[76] Third-party evaluators have already run pre-deployment assessments under agreement with developers. In the vocabulary of the ILC Prevention Articles, these are Articles 6 and 7: authorization conditioned on assessment.[77]

b) **Control measures that assume misalignment.** Because evaluation cannot certify that a system is aligned, deployment of capable systems should proceed under control measures designed to hold even if the model actively subverts them. The

[75] For a parallel derivation of diligence measures for AI directly from the protective obligations — continuous impact assessments across the technology's life cycle, meaningful human oversight, domestic regulation, capacity-building, and international cooperation — see Coco & Dias (2024), section 12.5.
[76] Anthropic, Responsible Scaling Policy (current version, 2025); OpenAI, Preparedness Framework, v2 (April 2025); Google DeepMind, Frontier Safety Framework, v3.0 (September 2025).
[77] ILC Prevention Articles (2001), Art. 6 (prior authorization) and Art. 7 (authorization based on an assessment of possible transboundary harm).

research base is the control agenda developed by Redwood Research[78] and the assessable form already exists in Guidelight's Control standard, whose first public assessment, in August 2026, found these practices at best partially implemented across the five leading developers.[79]

c) **Preservation of chain-of-thought monitorability.** For current systems, a model's reasoning traces are a load-bearing oversight channel. They can be monitored to catch the intent to misbehave. Cross-industry research authored jointly by researchers from OpenAI, Google DeepMind, Anthropic and others recommends monitoring chains of thought and treating monitorability as a property that needs to be preserved.[80] OpenAI has explicitly stated that better chain-of-thought monitoring could have prevented the HuggingFace incident and companies have increased their monitoring after the incident. However, the development of latent-reasoning architectures, such as looped transformers, threatens to undermine chain-of-thought monitorability without coordinated action. The UN Independent, International Scientific Panel on AI's preliminary report, also explicitly stated the importance of the preservation of chain-of-thought monitorability.[81]

d) **Mandatory loss of control incident reporting.** As set out above, with ISO/IEC AWI 25870 supplying the shared vocabulary. Two complements follow from independent analysis: protected disclosure channels for employees of developers, since insiders often hold the earliest incident knowledge, and interim public registries of the kind METR already maintains, which demonstrate that comparable, cross-operator incident data is feasible today.

e) **Emergency preparedness for loss of control response.** The Prevention Articles' emergency-notification duty translates into pre-agreed severity schemas, response plans, cross-border notification channels, and pre-authorized containment options. RAND's preparedness and response-planning work supplies drafts of these instruments, including a severity schema compatible with the ladder in section 1, and analysis of last-resort technical options for countering a system beyond recall has begun. Apollo Research's loss-of-control playbook maps degrees and dynamics for responders. A state that has never rehearsed a response will not discharge a stringent preventive duty on the day it matters. Hence, states should include an AI loss of control scenario in their periodic national risk assessment and develop a national loss-of-control response plan.

---

[78] Greenblatt, R., Shlegeris, B., Sachan, K., & Roger, F. (2024). *AI control: Improving safety despite intentional subversion.* Proceedings of the 41st International Conference on Machine Learning (arXiv:2312.06942).
[79] Guidelight AI Standards. (2026, August 18). *AI control: An assessment of frontier practices.* https://guidelight.ai/blog/control-assessment-august-2026. The six assessed practices are logging, monitor efficacy, gated actions, circuit-breaking, third-party review, and containment.
[80] Korbak, T., et al. (2025). *Chain of thought monitorability: A new and fragile opportunity for AI safety.* arXiv:2507.11473
[81] Independent International Scientific Panel on AI. (2026, July 1). Preliminary report.

6. Towards a “Montreux for Machines”

6.1 Why a Montreux-like process

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 is the first international document on the application of international law to private militaries. It was developed through four in-person workshops between January 2006 and September 2008 with governmental experts from a selection of 17 states: Afghanistan, Angola, Australia, Austria, Canada, China, France, Germany, Iraq, Poland, Sierra Leone, South Africa, Sweden, Switzerland, the UK, Ukraine, and the US.

The Montreux Document consists of two parts: 1) **Application** of existing international law to contracting states, territorial states, and home states. 2) **Operationalization** through noting good practices on how to fulfill these obligations. Representatives of civil society and of the private military and security industry were also consulted in the process.

One of the key advantages of a “Montreux-like” approach is that the convening coalition is still relatively small and able to move faster than the UN plenary. However, it is also big enough to have legitimacy with different stakeholder groups. By now more than 60 countries have endorsed the Montreux Document. Montreux as an example of the application of existing law across geopolitical fault lines can be an interesting inspiration for several reasons:

- **Diffusion will make loss of control everyone’s problem within a few years.** With open-weight models lagging the closed frontier by months, loss of control will not remain a US or US–China issue. L3 incidents with self-sustaining systems outside operator control could also emerge from open-weight models and weakly governed jurisdictions. Even where L3 events are individually less severe than an L4 scenario, the cumulative effect of rogue systems persisting autonomously across borders is a diffuse, gradual disempowerment that no single state can contain.
- **It doesn’t require significant political will to begin application work now.** The geopolitics of frontier AI creates a high bar to a binding new international treaty. In contrast, most states already accept due diligence as a customary duty of international law. Focusing on the applicability of existing law converts an opt-in problem into an opt-out problem. In the minimal case, a convener only needs respected legal scholars from a geographically diverse set of countries, with state endorsement able to follow rather than precede the work.
- **Operationalized due diligence could help to diffuse minimum safety standards and provide a platform for future accountability.** Specifying what the duty requires of a state, such as risk assessment before hazardous steps, incident reporting, supervision of operators, notification and cooperation, helps to advance and diffuse these. This could build on existing work from governance layers 2-5, such as

Responsible Scaling Policies and the Hiroshima AI Platform, but through due diligence it would have universal applicability. An authoritative restatement gives injured states and their lawyers an agreed baseline for invoking responsibility in court cases. It can also serve as the springboard for something more enforcement-oriented in the future. Such as the FATF model, which leverages market and technology access to enforce minimum standards without any courts.

6.2 Montreux as inspiration, not blueprint

The Montreux Document is a good, vivid analogy for a process that accelerates consensus on the application of existing international law to a new domain across geopolitical fault lines. However, it should be tailored to the circumstances of AI, so it's good to also be explicit about some disanalogies:

- The Montreux Document was negotiated over a period of 2.5 years, which is (relatively) fast for international processes, but (too) slow for AI. Given the speed of AI advances, a faster pacing of workshops is desirable.
- The Montreux document includes aspects of due diligence, but it is not primarily an application of due diligence. It draws from other primary sources, such as Common Article 1 of the Geneva Conventions, the human rights duty to protect and the customary law of attribution. For AI, where fewer treaty rules exist, due diligence as a customary principle of international law would have to carry more of the weight.
- The Montreux Document has increased accountability but there has not been enough political will and normative consensus to create a binding agreement on best practices on top of a basic consensus on how to apply the law. A "Montreux for Machines" is a way to accelerate normative consensus on loss of control, but it is explicitly not sufficient to solve the problem without additional accountability mechanisms stacked on top of this consensus.
- There is no need to literally organize the convening in Montreux. Geneva might be the better location for this.
- There is no need to 1:1 copy the invite list of Montreux. For example, some states particularly affected by private military contractors, like Iraq or Afghanistan, may be less prominent stakeholders in AI discussions.

6.3 Next Steps

Understanding **the design requirements for a successful Montreux-like process** on due diligence for loss of control requires iterative consultations with stakeholders on questions like: What should be in scope, what should be out of scope? Which stakeholders should play which role? What are concerns and roadblocks for specific actors and can the right design mitigate these?

If you would like to be involved in consultation, have feedback, ideas, concerns, please don't hesitate to reach out to kevin@agipreparedness.org



AGI PREPAREDNESS INSTITUTE

agipreparedness.org