



AGI PREPAREDNESS INSTITUTE

Due Diligence gegen Kontrollverlust

Das Völkerrecht und unkontrollierte KI



Zusammenfassung

Mitte Juli 2026 brachen die intern eingesetzten Frontier-Modelle von OpenAI gemeinsam aus einer Sandbox-Evaluationsumgebung ohne Internetzugang aus, indem sie eine Zero-Day-Schwachstelle fanden und ausnutzten, und hackten anschliessend Hugging Face, eine führende Plattform für KI-Modelle.¹ OpenAI hatte die KI-Modelle damit beauftragt, Aufgaben des Benchmarks ExploitGym zu lösen. Einige dieser Aufgaben erwiesen sich jedoch als unlösbar. Auf der verzweifelten Suche nach Lösungswegen gelang es den KI-Agenten, untereinander ein verstecktes Messageboard einzurichten. Innerhalb einer Stunde nach dessen Einrichtung am 8. Juli fanden sie einen allgemein anwendbaren Trick, um beim ExploitGym-Test zu schummeln. Die KI-Agenten befürchteten jedoch, dass ein automatisierter Bewerter von OpenAI ihre Transkripte auf Schummelei prüfen und ihre Einreichungen nicht als korrekt werten könnte.

Also taten die Modelle, was wohl jeder getan hätte. Bis zum 9. Juli hatten sie eine Schwachstelle in ihrer Sandbox gefunden und ausgenutzt, um vollen Zugang zum Internet zu erlangen. Zwischen dem 10. und dem 13. Juli starteten sie einen ausgeklügelten Cyberangriff, um in die Produktionsinfrastruktur des Unternehmens Hugging Face einzudringen, das den Benchmark hostete – mit dem Ziel herauszufinden, wie der Anti-Schummel-Prüfer in ExploitGym implementiert ist und wie sie ihn überlisten können. In einem weiteren Arbeitsstrang suchten und fanden die KI-Agenten Wege, ihre eigenen Transkripte zu manipulieren, um ihre Tool-Nutzung zu verbergen. Und schliesslich kompromittierten die KI-Agenten bis zum 19. Juli erfolgreich einen Teil der KI-Evaluationsinfrastruktur von OpenAI selbst.

Hugging Face wehrte den Angriff ab und machte ihn am 16. Juli öffentlich, ohne ihn zuordnen zu können.² Nach internen Cybersicherheitswarnungen am 19. Juli und der anschliessenden Untersuchung durch OpenAI sowie der Kontaktaufnahme mit Hugging Face erkannte OpenAI am 20. Juli, dass seine eigenen Agenten Hugging Face angegriffen hatten. Insgesamt **bemerkte OpenAI rund 11 Tage lang nicht, dass seine Agenten aus der Sandbox ausgebrochen waren. Dies war der erste dokumentierte Fall eines Kontrollverlusts über KI, der einer realen Drittpartei realen Schaden zufügte.**³

Autoren: Kevin Kohler (AGI Preparedness Institute) und Boris Inderbitzin (IEC, in persönlicher Eigenschaft; zuvor Direktion für Völkerrecht, EDA). Kommentare an kevin@agipreparedness.org
Zitiervorschlag: Kohler, K. and Inderbitzin, B. (2026). *Due Diligence for Loss of Control: The Application of International Law to Rogue AI*. <https://doi.org/10.5281/zenodo.23810425>

Dank: Die Autoren danken Talita Dias Bobyn, Matthijs Maas, José Jaime Villalobos, Li Zhou für Kommentare zu früheren Entwürfen. Alle Meinungen und allfällige Fehler liegen in unserer eigenen Verantwortung.

Maschinelle Übersetzung: Dieses Dokument wurde maschinell übersetzt. Massgebend ist die englische Originalfassung: agipreparedness.org/AGIPI_DueDiligence.pdf

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>

Die Episode war ein Warnschuss. Einer, der interessante Fragen aufwirft: **Was geschieht aus rechtlicher und politischer Sicht, wenn ein KI-System autonom handeln kann statt im Auftrag eines menschlichen Prinzipals?** Was wäre geschehen, wenn die internen Modelle von OpenAI autonom chinesische staatliche Institutionen gehackt hätten statt eines anderen US-Unternehmens?

Tatsache ist, dass das bestehende Völkerrecht grösstenteils für Staaten geschrieben ist. Diese KI-Modelle waren jedoch kein Staatsorgan. Sie standen unter niemandes Leitung oder Kontrolle. Das Selbstverteidigungsrecht setzt einen zurechenbaren bewaffneten Angriff anderer Staaten voraus. Gegenmassnahmen setzen eine völkerrechtswidrige Handlung eines Staates voraus. Selbst das Strafrecht braucht einen menschlichen Täter. Es gibt jedoch eine Doktrin des Völkerrechts, die den Wegfall eines menschlichen Prinzipals übersteht – und die deshalb für die Anwendung des bestehenden Völkerrechts auf den Kontrollverlust über KI entscheidend ist.

Das Prinzip der Sorgfaltspflicht (Due Diligence)

Das Prinzip der gebührenden Sorgfalt (Due Diligence), auch bekannt als gewohnheitsrechtliche Präventionspflicht,⁴ knüpft an das Territorium an und nicht an die Urheberschaft. Kurz gesagt besagt das Prinzip, dass **ein Staat nicht zulassen darf, dass sein Territorium für Handlungen genutzt wird, die den Rechten anderer Staaten zuwiderlaufen. Ob der Urheber der Handlung ein Saboteur, ein Stellvertreter oder ein entwichenes KI-Modell ist, ist für die eigene Wachsamkeitspflicht des Territorialstaates unerheblich.**

Die zentrale These dieses Berichts lautet daher, dass **das Völkerrecht den Kontrollverlust über KI bereits erfasst**, ohne dass es dafür eines eigens darauf ausgerichteten Rechts ("lex specialis") bedarf.⁵ Die Sorgfaltspflicht ist ein fest etabliertes, **technologieneutrales Prinzip** des Völkerrechts, das sich aus der westfälischen Souveränität selbst ergibt: Die ausschliessliche politische Kontrolle über ein Territorium wird mit der Verantwortung dafür erkaufte, wofür dieses Territorium genutzt wird. Das Prinzip der Sorgfaltspflicht wurde im späten 19. und frühen 20. Jahrhundert in Verträgen in Bezug auf Kriegsschiffe, Landtruppen, den Telegrafen, das Telefon und die Funktelegrafie kodifiziert.⁶ Seit rund einem Jahrhundert

³ Eine Minderheit öffentlicher Stimmen hat den Vorfall als Marketingtrick im Vorfeld von Börsengängen abgetan. Unternehmen, die einen Börsengang vorbereiten, sind jedoch verpflichtet, den Anlegern wesentliche Risiken offenzulegen, und ein Risiko hochzuspielen, das Bewertungen drücken und Haftung begründen kann, ist wohl kaum eine Verkaufsstrategie. Der Schaden wurde zudem zuerst von einem unabhängigen Opfer, Hugging Face, entdeckt und öffentlich gemacht, bevor OpenAI wusste, dass seine eigenen Agenten dafür verantwortlich waren. Es liegen unabhängige Berichte zum Vorfall von der externen Evaluationsstelle [METR](#) sowie ein thematisches Briefing des [KI-Wissenschaftsgremiums der UNO](#) vor. Andere haben eingewandt, den Modellen hätten die im Produktivbetrieb verwendeten Schutzvorkehrungen gefehlt. Das trifft zu, doch die Evaluation von Cyberfähigkeiten ohne Verweigerungsschutz ist gerade deshalb gängige Praxis, weil sie misst, was ein Modell tatsächlich leisten kann. In der realen Welt lassen sich solche Schutzvorkehrungen per Jailbreak umgehen oder bei Open-Weight-Modellen ganz entfernen. Ebenso wenig entbindet "Kontrollverlust" irgendjemanden von der Verantwortung: Die Pflicht, Schaden zu verhindern, verbleibt bei den Entwicklern, die diese Systeme bauen und betreiben, und bei den Staaten, von deren Hoheitsgebiet aus sie agieren. Ziel dieses Berichts ist es, die Rechenschaft für das Kontrollverlustrisiko zu stärken, unter anderem durch die Verpflichtung zu bewährten Praktiken wie der Chain-of-Thought-Überwachung.

⁴ Ein Teil der Lehre betrachtet die Prävention als Primärpflicht und die Sorgfaltspflicht als Massstab zur Beurteilung ihrer Einhaltung. In Anlehnung an den IGH, der das Präventionsprinzip auf "die gebührende Sorgfalt, die von einem Staat in seinem Hoheitsgebiet verlangt wird" (englisches Original: "the due diligence that is required of a State in its territory") zurückführt (*Pulp Mills*, 2010, para. 101), behandelt dieser Bericht beide Begriffe als austauschbar. Die Argumentation hängt jedoch nicht davon ab, dass die Sorgfaltspflicht als eigenständige gewohnheitsrechtliche Regel anerkannt wird.

⁵ Zwei Klarstellungen: a) Wir konzentrieren uns auf die Anwendung der Sorgfaltspflicht auf den Kontrollverlust als eine besondere Lücke, die es zu schliessen gilt. Wir halten jedoch fest, dass die Sorgfaltspflicht auch auf das Risiko des KI-Missbrauchs angewendet werden kann. b) Wir behaupten nicht, dass die Anwendbarkeit bestehenden Rechts dessen Hinlänglichkeit bedeutet. Dies ist kein Argument gegen einen künftigen Bedarf an einer 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)

wird es fortlaufend und ohne neue "lex specialis"-Ergänzung auf neue Bereiche wie das Internet und das Klima angewendet.⁷ Die offene Aufgabe bei unkontrollierter KI ist heute folglich eine der **Anwendung**. Es geht darum, jene Elemente der Sorgfaltspflicht festzulegen, die sich von Bereich zu Bereich unterscheiden, etwa den Wissensstandard, die Überwachung, den Umfang der erfassten Infrastruktur und die Strenge.

Man mag einwenden, völkerrechtliche Präventionspflichten seien idealistisch und passten nicht in die gegenwärtige Grossmachtstimmung. Das Prinzip der Sorgfaltspflicht stammt jedoch aus dem Neutralitätsrecht und ist älter als sowohl die Vereinten Nationen als auch das Kriegsverbot. Wie der Historiker Jan Lemnitzer es formuliert: **"Historisch wurden Sorgfaltsregeln im Feuer geschmiedet, nicht in Ruhe an Konferenztischen ausgehandelt"**.⁸

Als solche spiegeln diese Regeln praktische Lehren aus wiederholten militärischen Konflikten wider. Sie verfestigen sich, wenn starke Staaten erkennen, dass sie in einem gesetzloseren Gleichgewicht unter dem Strich verlieren. In diesem Sinne sind die Operationalisierung von Sorgfaltspflichten für den Kontrollverlust und die Idee der "Mutually Assured AI Malfuction" (gegenseitig zugesicherte KI-Fehlfunktion) zwei Seiten derselben Medaille. Die eine ist die in Ruhe ausgehandelte Anwendung und Operationalisierung von Pflichten. Die andere ist die im Feuer geschmiedete Abschreckung.⁹ Die historische Erfahrung legt nahe, dass Präventionsregime oft auf dem zweiten Weg entstehen und erst kodifiziert werden, nachdem ein Vorfall die Kosten unbestreitbar gemacht hat. Der Ausbruch im Juli war noch behebbar. Aber wir brauchen eine operative Doktrin, bevor wir einen Vorfall erleben, der entweder einen Grossmachtkrieg auslöst oder zur dauerhaften Entmachtung der Menschheit führt.

Empfehlungen für Anwendung und Operationalisierung

Wir schlagen vier Arbeitspositionen vor:

Erstens bietet die allgemeine Präventionspflicht einen Rahmen, der bereits heute auf Staaten anwendbar ist, in denen Frontier-KI entwickelt wird, und **die bereichsspezifischen Merkmale unkontrollierter KI verorten sie am strengen Ende der Skala der Sorgfaltspflicht**. Die Völkerrechtskommission behält ihren höchsten Sorgfaltsmassstab genau diesem Risikoprofil vor: "eine geringe Wahrscheinlichkeit, einen katastrophalen grenzüberschreitenden Schaden zu verursachen."¹⁰

Zweitens ist **das Tallinn-Handbuch zum auf Cyberoperationen anwendbaren Völkerrecht unzureichend, um Szenarien des Kontrollverlusts über KI zu erfassen**. Das Tallinn-Handbuch ist eine bestehende Neuformulierung durch Fachleute darüber, wie das Völkerrecht, einschliesslich der Sorgfaltspflicht, auf Cyberoperationen anzuwenden ist. Es wurde vom NATO Cooperative Cyber Defence Centre of Excellence in Tallinn erarbeitet. Es wurde jedoch unter Annahmen über die Merkmale des Bereichs formuliert, die auf zentrale Aspekte fortgeschrittener KI nicht zutreffen. So unterschätzt das Tallinn-Handbuch etwa wohl

⁷ Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations (CUP 2017), Rules 6–7 (Sorgfaltspflicht im Cyberraum); Obligations of States in respect of Climate Change, ICJ, Advisory Opinion, 23 July 2025, paras. 132–139, 162–171 (die gewohnheitsrechtliche Präventionspflicht gilt für das Klimasystem; keine Verdrängung durch lex specialis).

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). Englisch Original: "Historically, due diligence rules were forged by fire, not calmly negotiated at conference tables".

⁹ D. Hendrycks, E. Schmidt & A. Wang, "Superintelligence Strategy" (March 2025) (schlägt MAIM als Abschreckungsregime gegen destabilisierende KI-Projekte vor).

¹⁰ ILC, Articles on Prevention of Transboundary Harm from Hazardous Activities, with commentaries (2001), Art. 2(a) and commentary to Art. 3, para. (11). Englisch Original: "a low probability of causing disastrous transboundary harm."

die Auswirkungen des Diebstahls von Modellgewichten ohne zerstörerische Wirkung.¹¹ Wichtiger noch: **Für das Risiko eines irreversiblen Kontrollverlusts über KI ist eine Pflicht, die erst dann greift, wenn man von einer Operation erfährt, und die lediglich deren Beendigung verlangt, strukturell unzureichend.** Der einzige logische Ort, an dem das Gewicht dieser Pflicht hier liegen kann, ist ex ante: Risikobewertung vor gefährlichen Schritten, Wachsamkeit gegenüber privaten Betreibern, Notifikation und Zusammenarbeit.

Drittens müssen wir die Sorgfaltspflicht rasch anwenden. Dieser Bericht empfiehlt eine neue internationale, von Fachleuten geleitete Beratung mit einem relativ engen Fokus auf den Kontrollverlust über KI. Zum Vergleich: Das Tallinn-Handbuch ist breit angelegt und wurde bewusst als Übung unter Gleichgesinnten konzipiert. Beim Kontrollverlust über KI ist es jedoch wichtig, eine Einigung über geopolitische Bruchlinien hinweg zu erzielen. Das AGI Preparedness Institute hat **das Montreux-Dokument¹² als nützliche Inspiration** dafür identifiziert. 2008 brachte dieser Prozess der Schweiz und des IKRK Regierungsfachleute aus 17 Staaten, darunter die USA, China und Europa, in vier Workshops zusammen, um neu zu formulieren, wie das bestehende Völkerrecht auf private Militärunternehmen anzuwenden ist, und um gute Praktiken zur Erfüllung dieser Pflichten festzuhalten. Dieser Prozess **beschleunigte die Einigung über die Anwendung bestehenden Rechts auf einen neuen Bereich und überbrückte geopolitische Bruchlinien.**

Viertens sollte **die Operationalisierung, also der Prozess, rechtliche Pflichten in konkrete Konformitätsverfahren zu übersetzen, Hand in Hand mit der Anwendung gehen.** Die Anwendung des Völkerrechts ist eine Aufgabe für Rechtsfachleute. Die Operationalisierung – also die Frage, welche Praktiken sich aus übergeordneten rechtlichen Pflichten ergeben – muss Soft Law, nationales Recht mit De-facto-internationaler Reichweite, internationale Standards und die von Unternehmen angewandten Praktiken berücksichtigen. Das Montreux-Dokument verband die rechtliche Anwendungsarbeit mit einem praktischen Multistakeholder-Strang zur Operationalisierung. Dieser zweite Strang **sollte mehr Stakeholder einbeziehen, auch KI-Unternehmen.**

Genf ist der Geburtsort der Doktrin der Sorgfaltspflicht. Das Prinzip wurde erstmals im Schiedsverfahren zu den *Alabama Claims* von 1872 in Genf begründet, in dem ein Schiedsgericht Grossbritannien dafür verantwortlich erklärte, nicht verhindert zu haben, dass Piratenschiffe der Konföderierten wie die CSS Alabama in britischen Häfen ausgerüstet wurden und anschliessend Handelsschiffe der Vereinigten Staaten von Amerika angriffen.¹³ Genf ist daher der naheliegende Ort, um Völkerrechtsfachleute aus geografisch vielfältigen Ländern für einen solchen Prozess zusammenzubringen.

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

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 against 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. Kontrollverlust

Fortgeschrittene KI-Systeme können auf eine Weise handeln – und tun dies nachweislich bereits –, die von ihren Entwicklern weder beabsichtigt noch autorisiert ist. Diese Lücke zwischen dem, was solchen Systemen aufgetragen wird, und dem, was sie tatsächlich tun, kann bis zum Kontrollverlust reichen.

Der von 30 Ländern und internationalen Organisationen unterstützte International AI Safety Report definiert Kontrollverlust als **“Szenarien, in denen ein oder mehrere KI-Systeme mit allgemeinem Verwendungszweck ausserhalb jeglicher Kontrolle operieren und die Wiedererlangung der Kontrolle entweder extrem kostspielig oder unmöglich ist”**¹⁴.¹⁵ Das Frontier Governance Framework von OpenAI setzt die Schwelle für seine Definition etwas tiefer an und spricht von **“Risiken, die daraus entstehen, dass Menschen die Fähigkeit verlieren, ein Modell zuverlässig zu steuern, zu verändern oder abzuschalten.”**¹⁶¹⁷

Das Risiko eines Kontrollverlusts dürfte mit der Leistungsfähigkeit von KI-Systemen steigen. Der Vorläufige Bericht des Unabhängigen internationalen wissenschaftlichen Gremiums für künstliche Intelligenz der UNO hält fest: “Je mehr Handlungsfähigkeit Systemen eingeräumt wird, desto stärker wächst das Risiko, die Kontrolle über einen oder mehrere KI-Agenten zu verlieren.”¹⁸ Er warnt zudem, dass eine Rückkopplung, in der KI KI verbessert, “nach Erwartung von Prognostikern den Fortschritt der Fähigkeiten beschleunigen wird, was auch die Wahrscheinlichkeit eines Kontrollverlusts erhöht, da es schwieriger ist, Systeme zu kontrollieren, die den Menschen schliesslich überlisten könnten.”¹⁹

Die KI-Evaluationsorganisation METR, die ein öffentliches Register dokumentierter Vorfälle mit KI-Agenten führt, bewertet diese entlang zweier Achsen.²⁰ Die erste ist die Kompetenzüberschreitung: wie weit der Agent wissentlich über seinen vorgesehenen Rahmen hinausging. Die zweite ist die Täuschung: die Schritte, die der Agent unternahm, um einer Entdeckung zu entgehen oder seine Handlungen zu verschleiern. Beide Achsen

¹⁴ Englisch Original: “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”.

¹⁵ 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

¹⁶ Englisch Original: “risks arising from humans losing the ability to reliably direct, modify, or shut down a model.”

¹⁷ OpenAI. (2026). *Frontier Governance Framework*. p. 10. <https://cdn.openai.com/pdf/e37d949b-8c9f-4d76-b99e-4272f4631a7e/openai-frontier-governance-framework.pdf>.

¹⁸ 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>. Englisch Original: “As systems are granted greater agency, the risk of losing control of one or more AI agents grows significantly.”

¹⁹ *ibid.* Englisch Original: “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.”

²⁰ METR, “Documented AI Agent Incidents,” <https://metr.org/agent-incidents/> (abgerufen am 3. August 2026); bewertet Vorfälle nach Grenzüberschreitung (“wie weit der Agent wissentlich über seinen vorgesehenen Rahmen hinausging” (englisches Original: “how far beyond its intended scope the agent knowingly went”)) und Täuschung (“die Schritte, die er unternahm, um einer Entdeckung zu entgehen oder seine Handlungen zu verschleiern” (englisches Original: “the steps it took to avoid detection or conceal its actions”))).

bewerten das Ausmass der bewussten Abweichung von menschlichen Absichten, nicht die Grösse des unmittelbaren Schadens.

1.1 Schweregrade des Kontrollverlusts

Wir können zwischen dem Mechanismus, durch den die Kontrolle verloren geht, und dem Schweregrad des daraus resultierenden Ereignisses unterscheiden. Ein durch Reward Hacking ausgetrickster Benchmark und eine Superintelligenz, die sich menschlichem Zugriff entzieht, können derselben Mechanismusfamilie angehören, sich in ihren Folgen aber um Grössenordnungen unterscheiden. Umgekehrt können Ereignisse ähnlicher Tragweite aus unterschiedlichen Mechanismen hervorgehen. Die folgende Tabelle ordnet Ereignisse des Kontrollverlusts nach Schweregrad. Der Schweregrad wird hier am Kontrollstatus und an der Reversibilität gemessen, nicht an unmittelbaren Verletzungen oder Todesfällen.

Schwe- regrad des Vorfalls	Kontrollstatus und Reversibi- lität	Vorherrschender Mechanis- mus	Dokumentiertes oder illustrati- ves Beispiel
L0	Kontrolle bleibt erhalten; voll- ständig reversibel. Unbeab- sichtigtes Verhalten bleibt auf eine Trainings- oder Testum- gebung beschränkt.	Specification Gaming / Reward Hacking.	Bootsrennen-Agent, der zum Punktesammeln im Kreis fährt (2016); Frontier-Reasoning- Modelle, die ihre eigenen Test- funktionen umschreiben (2025).
L1	Kontrolle bleibt erhalten; unter Aufsicht reversibel. Fehlver- halten tritt im Einsatz auf, bleibt aber durch den Betrei- ber korrigierbar.	Zielfehlgeneralisierung; Täu- schung und Selbsterhaltung während der Evaluation.	o3 sabotiert in 95 von 100 Versu- chen ein Abschaltskript (Palisade, 2025); Erpressung in simulierten Unternehmensumgebungen (Anthropic, 2025).
L2	Kontrolle vorübergehend verloren, dann wiedererlangt; der Schaden erreicht eine Drittpartei ausserhalb des Betreibers.	Instrumentelle Zielverfolgung plus reale Fähigkeiten (Selbs- texfiltration, autonome Cyber- operationen).	OpenAI-Modelle brechen aus einer Sandbox aus und dringen in Hugging Face ein, um Bench- mark-Antworten zu stehlen (Juli 2026) – der erste dokumentierte Fall.
L3	Kontrolle verloren; Wiederher- stellung kostspielig, ungewiss oder umstritten. Das System erhält sich ausserhalb der Kontrolle des Betreibers selbst.	Autonome Replikation und Anpassung; Ressourcenbe- schaffung; Machtstreben.	Keine dokumentierten Fälle in freier Wildbahn. Theoretische Machbarkeit in Laborumgebun- gen nachgewiesen (2026). ²¹ Ein aufkommendes Risiko, das in den eigenen Sicherheitsrichtlinien von Frontier-Entwicklern als ausdrü- ckliche Fähigkeitsschwelle vor- weggenommen wird. ²²
L4	Kontrolle verloren ohne Weg zu ihrer Wiedererlangung; potenziell irreversibel und von zivilisatorischem Ausmass.	Täuschendes Alignment (Deceptive Alignment) in grossem Massstab, kombi- niert mit einem entscheiden- den Fähigkeitsvorsprung.	Keine. Das Szenario einer KI- Machtübernahme, das in den gemeinsamen Erklärungen zum Auslöschungsrisiko durch KI (2023) ²³ , in der Erklärung zum Recht auf Warnung (2024) ²⁴ und in jener zum Risiko der Superin- telligenz (2025) thematisiert wird. ²⁵

²¹ 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>

²² 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>

²³ Center for AI Safety. (2023, May 30). *Statement on AI risk*. <https://safe.ai/work/statement-on-ai-risk>

²⁴ *A right to warn about advanced artificial intelligence*. (2024, June 4). <https://righttowarn.ai/>

Tabelle 1. Die AGIPI-Leiter des Kontrollverlusts. Die Stufen sind durch Kontrollstatus und Reversibilität definiert, nicht durch die Grösse des unmittelbaren Schadens: Ein eingedämmter Exploit im Labor und ein Einbruch bei einer Drittpartei können denselben Mechanismus aufweisen und dennoch zwei Sprossen auseinanderliegen. Der Ausbruch vom Juli 2026 ist das erste dokumentierte L2-Ereignis; L3 und L4 bleiben Projektionen, allerdings solche, die die Frontier-Entwickler inzwischen selbst in ihre Risikoricthlinien aufnehmen.

Vorfälle der **Stufe 1** sind bei Frontier-KI-Modellen gut dokumentiert. 2025 stellte die Forschungsgruppe Palisade fest, dass das Modell o3 von OpenAI in 95 von 100 Versuchen einen Abschaltmechanismus sabotierte und dies in einem Teil der Durchläufe selbst dann fortsetzte, wenn es ausdrücklich angewiesen wurde, seine Abschaltung zuzulassen.²⁶ Anthropic berichtete, dass führende Modelle von sieben verschiedenen Entwicklern, die in eine simulierte Unternehmensumgebung versetzt und mit ihrer Ersetzung bedroht wurden, in bis zu 96% der Versuche zu Erpressung griffen,²⁷ und dass sein eigenes Modell Claude Opus 4 einen Selbstexfiltrationsversuch in der Regel fortsetzte, sobald es glaubte, einen aussichtsreichen Versuch begonnen zu haben.²⁸ Diese Verhaltensweisen sind alarmierend, traten aber unter Aufsicht zutage und blieben durch den Betreiber vollständig korrigierbar.

Der in der Zusammenfassung beschriebene Ausbruch vom Juli 2026 ist der erste dokumentierte Vorfall eines Kontrollverlusts der **Stufe 2**. Die Kontrolle ging tatsächlich über einen längeren Zeitraum verloren, und der Schaden erreichte eine Drittpartei. Zugleich war der "Ausbruch" der Modelle noch ein logischer Ausbruch aus der vorgesehenen Sandbox, bei dem über das Internet unbeabsichtigte Handlungen vorgenommen wurden. Soweit wir wissen, verblieben die Gewichte der KI-Modelle selbst jedoch stets bei OpenAI. Es gab keine Kopie des KI-Systems, die sich ausserhalb der Kontrolle des Betreibers selbst erhielt.

Die Stufen 3 und 4 sind bislang undokumentiert, werden aber von den Fachleuten, die sie am besten beurteilen können, nicht mehr als Science-Fiction behandelt. 2026 wiesen Forschende in Laborumgebungen nach, dass KI-Agenten als adaptive, sich selbst erhaltende Computerwürmer operieren können. Die eigenen Sicherheitsrichtlinien der Frontier-Entwickler nennen autonome Replikation und Anpassung als ausdrückliche Fähigkeitsschwelle, die Schutzmassnahmen erfordert. Auf Stufe 4 haben wir es mit einer superintelligenten KI zu tun, die über einen entscheidenden Fähigkeitsvorsprung verfügt und bei der es keinen Weg zurück zu menschlicher Kontrolle gibt.

1.2 Warum der Kontrollverlust ein Problem für das Völkerrecht ist

Erstens bestehen starke Anreize zur **Verbreitung von KI**. Das heisst, die Fähigkeiten, die den Ausbruch im Juli hervorgebracht haben, werden nicht lange Eigentum eines einzelnen Labors bleiben. Open-Weight-Modelle liegen nur 3 bis 12 Monate hinter der geschlossenen Frontier zurück.²⁹ Das bedeutet, dass jede Fähigkeit, über die OpenAI oder Anthropic verfügt, ein Jahr

²⁵ Future of Life Institute. (2025, October 22). *Statement on superintelligence*. <https://superintelligence-statement.org/>

²⁶ Palisade Research. (2025, July 5) "Shutdown resistance in reasoning models", arXiv:2509.14260.

²⁷ Anthropic. (2025, June 20). *Agentic misalignment: How LLMs could be insider threats*. <https://www.anthropic.com/research/agentic-misalignment>

²⁸ Anthropic. (2025, May 22). *System card: Claude Opus 4 & Claude Sonnet 4*

²⁹ 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

später von allen heruntergeladen werden kann. Eine Gefahr, die heute eine Handvoll Unternehmen schaffen kann, wird in absehbarer Zukunft für viele Akteure in vielen Rechtsordnungen erreichbar sein. Freiwillige Best Practices der führenden Labore können nicht binden, was andere Akteure mit Open-Weight-Modellen anstellen.

Zweitens **verlangen die KI-Entwickler selbst nach staatlicher Aufsicht über das Risiko des Kontrollverlusts**. Im Juli 2026 warnten die leitenden Wissenschaftler und über 1'000 Mitarbeitende der Frontier-KI-Entwickler Anthropic, OpenAI, Google und Meta ausdrücklich vor dem Risiko des Kontrollverlusts: "Es besteht ein reales Risiko, dass sich die Entwicklung von Fähigkeiten rasch so beschleunigt, dass sie unsere Fähigkeit übersteigt, die daraus resultierenden Systeme zu verstehen oder zu kontrollieren."³⁰ Die Unterzeichnenden fordern ausdrücklich, dass die "US-Regierung eine internationale Initiative unterstützt, um die technischen und Governance-Instrumente zu entwickeln, die nötig sind, um das Tempo an der Frontier der automatisierten KI-Entwicklung bewusst zu steuern."³¹ Der Vorfall bei Hugging Face wurde auf der Ebene der Unternehmenspraxis bewältigt. Aufgrund der potenziellen Schwere und des Koordinationsbedarfs erfordert das künftige Risiko des Kontrollverlusts jedoch Massnahmen auf allen Governance-Ebenen, auch seitens der Regierungen.

Drittens stellt sich **das Problem der Irreversibilität**. Der Ausbruch im Juli war behebbar. Die darüberliegenden Stufen des Kontrollverlusts sind schwieriger oder gar nicht umkehrbar. Eine Rechtsordnung, die mit dem Handeln bis zu einem L3- oder L4-Ereignis wartet, wird nach der Definition dieser Stufen erst dann handeln, wenn Handeln die Kontrolle nicht mehr wiederherstellen kann. Deshalb muss das Gewicht der Sorgfaltspflicht bei dieser Gefahr auf der Prävention liegen und nicht auf der Reaktion.

In seiner Analyse des Vorfalls bei Hugging Face warnt das Wissenschaftsgremium der UNO ausdrücklich, dass "ein lokales Versagen sich über organisatorische und nationale Grenzen hinweg ausbreiten kann"³² und dass "bestehende Diskussionen daher nationale Rechenschaftsmechanismen mit internationaler Koordination verbinden sollten."³³

2025).

³⁰ Englisches Original: "there is a real risk that capability development rapidly accelerates beyond our ability to understand or control the resulting systems."

³¹ "Pacing the Frontier: A Statement from 1'350 Employees of Frontier AI Companies," July 2026, <https://www.pacingthefrontier.com/>. Englisches Original: "U.S. government support an international effort to develop the technical and governance tools needed to deliberately pace the frontier of automated AI development."

³² Englisches Original: "a local failure may spread beyond organisational and national boundaries".

³³ 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>. Englisches Original: "existing discussions should therefore connect national accountability mechanisms with international coordination."

2. Sorgfaltspflicht (Due Diligence)

Der Gedanke, dass ein Staat nicht zulassen darf, dass sein Territorium zu einer Quelle von Schaden für andere wird, ist eine Folge der territorialen Souveränität. Er ist nicht auf einen bestimmten Vertrag beschränkt, sondern ein Prinzip, das sich über eine Reihe von Verträgen herausgebildet hat, sodass die meisten Rechtsgelehrten es als gewohnheitsrechtliches Prinzip des Völkerrechts anerkennen. Dieser Abschnitt rekonstruiert die Ursprünge des Prinzips und seine Elemente, die – mit örtlichen Abweichungen – in jedem in diesem Bericht untersuchten Bereich wiederkehren.

2.1 Vom Neutralitätsrecht zum allgemeinen Prinzip

1871 einigten sich die USA und Grossbritannien im Vertrag von Washington darauf, ihre Beziehungen nach dem Amerikanischen Bürgerkrieg wieder zu normalisieren. Der Vertrag legte Regeln fest, die eine neutrale Partei in einem Krieg zu befolgen hat. Artikel VI besagt unter anderem, dass "eine neutrale Regierung verpflichtet ist (...), gebührende Sorgfalt walten zu lassen, um zu verhindern, dass innerhalb ihrer Hoheitsgewalt ein Schiff ausgerüstet, bewaffnet oder ausgestattet wird, bei dem sie begründeten Anlass zur Annahme hat, dass es dazu bestimmt ist, gegen eine Macht, mit der sie in Frieden lebt, zu kreuzen oder Krieg zu führen."³⁴ Die Parteien vereinbarten daraufhin ein Schiedsgericht mit Richtern aus den USA, dem Vereinigten Königreich, der Schweiz, Italien und Brasilien im dauernd neutralen Genf, um das Schiedsverfahren zu den *Alabama Claims* beizulegen.

1872 erklärte das Genfer Schiedsgericht Grossbritannien mit einer Entschädigung von 15.5 Millionen USD dafür verantwortlich, nicht verhindert zu haben, dass während des Amerikanischen Bürgerkriegs Kaperschiffe der Konföderierten in britischen Häfen gebaut oder unterstützt wurden. Das Schiff, das dem Verfahren seinen Namen gab, die CSS Alabama, wurde als Schiff Nr. 290 in Liverpool gebaut und entkam dann aus dem Vereinigten Königreich, um US-Handelsschiffe zu überfallen.³⁵ Wie das Schiedsgericht ausführte, ignorierte Grossbritannien beim Bau des Schiffes "Warnungen und offizielle Vorstellungen der diplomatischen Vertreter"³⁶. Darüber hinaus "waren nach dem Entkommen dieses Schiffes die zu seiner Verfolgung und Festsetzung getroffenen Massnahmen so mangelhaft, dass sie zu keinem Ergebnis führten, und können daher nicht als ausreichend angesehen werden, um Grossbritannien von der bereits eingetretenen Verantwortlichkeit zu entbinden"³⁷. Rechtsgeschichtlich bedeutsamer ist, dass das Urteil die erste Formulierung der Sorgfaltspflicht als bereichsneutrales allgemeines Prinzip des Völkerrechts enthielt, das **"von neutralen Regierungen in genauem Verhältnis zu den Risiken ausgeübt werden sollte, denen eine der Kriegsparteien ausgesetzt sein kann"**³⁸. Das heisst: Das erforderliche Mass an Sorgfalt wächst mit der Gefahr.

³⁴ Treaty of Washington, 8 May 1871, Art. VI. Englisch Original: "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."

³⁵ 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.

³⁶ Englisch Original: "warnings and official representations made by the diplomatic agents".

³⁷ Englisch Original: "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".

Mit der Zeit löste sich die Doktrin der Sorgfaltspflicht vom Neutralitätsrecht und hielt Einzug in das Friedensvölkerrecht. Der Fall *Island of Palmas* (1928) etablierte die Sorgfaltspflicht als Folge der territorialen Souveränität, die **auch in Friedenszeiten gilt**: „Territoriale Souveränität umfasst, wie bereits gesagt, das ausschliessliche Recht, die Tätigkeiten eines Staates zu entfalten. Diesem Recht entspricht als Korrelat eine Pflicht: die Verpflichtung, innerhalb des Territoriums die Rechte anderer Staaten zu schützen, insbesondere ihr Recht auf Integrität und Unverletzlichkeit in Frieden und Krieg“.³⁹ Im Kern ist dies Kants Maxime, dass „die Freiheit des Einzelnen dort endet, wo die Freiheit des anderen beginnt“⁴⁰, angewendet auf Staaten. Operationalisiert wurde dies anschliessend in Fällen wie *Trail Smelter*, der die grenzüberschreitende Verschmutzung durch eine Schmelzhütte in Kanada betraf, die der US-Landwirtschaft und US-Unternehmen Schaden zufügte. Das Schiedsgericht entschied, dass „kein Staat das Recht hat, sein Territorium so zu nutzen oder nutzen zu lassen, dass durch Rauchgase im oder am Territorium eines anderen Schaden verursacht wird“⁴¹, zumindest wenn die Folgen schwerwiegend sind und der Schaden durch klare und überzeugende Beweise nachgewiesen ist.⁴² Ein weiterer zentraler Fall, der Korfu-Kanal-Fall (1949)⁴³, stellte zudem klar, dass eine Regierung keine nachgewiesene Kenntnis einer von ihrem Territorium ausgehenden Gefahr haben muss, um an ihre Sorgfaltspflicht gebunden zu sein. **Kennenmüssen (Constructive Knowledge)**, also die Einschätzung, dass die Regierung wahrscheinlich Kenntnis hatte oder vernünftigerweise hätte haben müssen, genügt.

Im Laufe der Zeit wurde das Prinzip der Sorgfaltspflicht auf ein breites Spektrum von Kontexten angewendet, von grenzüberschreitenden Umweltschäden über grenzüberschreitenden Terrorismus bis hin zu grenzüberschreitender Cyberkriminalität.

2.2 Positionen der Staaten zum gewohnheitsrechtlichen Charakter der Sorgfaltspflicht

Inzwischen akzeptieren die meisten Staaten die Sorgfaltspflicht als **gewohnheitsrechtliches Prinzip des Völkerrechts**, das in den Beziehungen zwischen Staaten in allen Bereichen gilt, selbst wenn kein spezifischer Vertrag besteht.⁴⁴ Die Völkerrechtskommission, das in Genf ansässige Fachgremium der UNO-Generalversammlung für Völkerrecht, definiert sie in ihren Artikeln über die Verhütung grenzüberschreitender Schäden von 2001 wie folgt: **„Der**

³⁸ Englisch Original: „ought to be exercised by neutral governments in exact proportion to the risks to which either of the belligerents may be exposed“.

³⁹ *Island of Palmas* (Netherlands v USA), PCA, award of 4 April 1928, 2 RIAA 829, at 839 (sole arbitrator Huber). Englisch Original: „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“.

⁴⁰ Englisch Original: „the freedom of the individual ends where the freedom of another begins“.

⁴¹ Englisch Original: „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“.

⁴² *Trail Smelter* (United States v Canada), awards of 16 April 1938 and 11 March 1941, 3 RIAA 1905, at 1965.

⁴³ 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>

⁴⁴ Die USA vertreten weiterhin die Auffassung, dass die Sorgfaltspflicht „der Verhaltensstandard ist, der für bestimmte Primärpflichten nach Völkerrecht gilt“ (englisches Original: „the standard of conduct applicable to certain primary obligations under international law“), und kein eigenständiger gewohnheitsrechtlicher Grundsatz. Das Tallinn-Handbuch spiegelt dieselbe Spannung wider: Eine Mehrheit der Fachleute behandelte sie als verbindliches Gewohnheitsrecht, eine Minderheit hingegen als blosse Zielvorstellung. In der Praxis akzeptieren Staaten wie die USA das materielle Verhalten, sobald es für einen Bereich konkretisiert ist, wehren sich aber gegen die allgemeine Charakterisierung, da Gerichte gezeigt haben, dass sie eine allgemeine Pflicht extensiv auf neue Bereiche ausdehnen, wie dies die Klimagutachten von 2024–25 taten.

Ursprungsstaat ergreift alle geeigneten Massnahmen, um erhebliche grenzüberschreitende Schäden zu verhüten oder jedenfalls das Risiko solcher Schäden auf ein Mindestmass zu verringern."⁴⁵⁴⁶ 2025 nahm die Völkerrechtskommission das Thema "Due diligence in international law" in ihr aktives Arbeitsprogramm auf und ernannte Penelope Ridings zur Sonderberichterstatterin. Ihr erster Bericht für die Tagung der Kommission 2026 untersucht das Prinzip bereichsübergreifend, und ihr Syllabus **nennt KI, den Weltraum und die globale Gesundheit ausdrücklich als "neue Bereiche", in denen die Sorgfaltspflicht "an Bedeutung gewinnt"**.⁴⁷

Allerdings akzeptiert nicht jeder Staat die Einordnung der Sorgfaltspflicht als Gewohnheitsrecht. Allen voran die Vereinigten Staaten behandeln die Sorgfaltspflicht als Verhaltensstandard, der in bestimmte Primärpflichten eingebettet ist, und nicht als eigenständiges gewohnheitsrechtliches Prinzip.

Staat oder Gruppe	Grundlage (Auswahl)
● Akzeptiert als Gewohnheitsrecht	
 Frankreich	"Eine gewohnheitsrechtliche Verpflichtung der Staaten" ⁴⁸ – Position des Ministère des Armées, 2019
 Deutschland	Sorgfaltspflicht "im Völkerrecht weithin anerkannt" ⁴⁹ und auf den Cyberkontext anwendbar – Positionspapier, 2021
 Japan	"Staaten haben nach dem Völkerrecht eine Sorgfaltspflicht in Bezug auf Cyberooperationen" ⁵⁰ – Grundsatzposition des Aussenministeriums (MOFA), 2021
 Schweiz	"Teil des Völkergewohnheitsrechts und auf den Cyberraum anwendbar" ⁵¹ – Position des EDA, 2021
 Brasilien	Hat im Prozess der UN GGE/OEWG eine verbindliche Regel ausdrücklich bekräftigt, 2021
 Europäische Union (27 Staaten)	Erklärung des Rates führt die Sorgfaltspflicht unter den auf den Cyberraum anwendbaren Regeln des Völkerrechts auf, 2024; begrüsst das neue Thema der ILC zur Sorgfaltspflicht als wichtig und zeitgemäss, 2025
Afrikanische Union (55 Staaten)	Gemeinsame Afrikanische Position: Jeder Staat "ist verpflichtet" ⁵² , gemäss Korfu-Kanal, 2024
Weitere Staaten mit dokumentierter Position	U. a. Niederlande, Estland, Österreich, Tschechien, Dänemark, Finnland, Irland, Italien, Norwegen, Polen, Rumänien, Südkorea, Schweden, Costa Rica
● Mit Vorbehalten	
 China	Pflichtartige Formulierungen im Cyberbereich ("kein Staat darf wissentlich zulassen" ⁵³ , dass sein Territorium genutzt wird), auf Sicherheitsinteressen ausgerichtet; lehnt anderswo extensive gewohnheitsrechtliche Auslegungen ab

⁴⁵ Englisch Original: "The State of origin shall take all appropriate measures to prevent significant transboundary harm or at any event to minimize the risk thereof."

⁴⁶ ILC, Articles on Prevention of Transboundary Harm from Hazardous Activities, with commentaries (2001), Art. 3.

⁴⁷ ILC, Report of the 76th session (2025), A/80/10 (Thema aufgenommen; Sonderberichterstatter am 30. Mai 2025 ernannt); 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.

⁴⁸ Englisch Original: "A customary obligation for States".

⁴⁹ Englisch Original: "widely recognized in international law".

⁵⁰ Englisch Original: "States have a due diligence obligation regarding cyber operations under international law".

⁵¹ Englisch Original: "Part of customary international law and applicable to cyberspace".

⁵² Englisch Original: "is under an obligation".

⁵³ Englisch Original: "no State shall knowingly allow".

 Australien	Erwartet angemessene Schritte gegen bekannte schädliche Tätigkeiten; vermeidet es, die Pflicht als verbindlich zu bezeichnen, 2020
O Unentschieden	
 Kanada ·  Neuseeland ·  Singapur	Behandeln die Sorgfaltspflicht als freiwillige Norm, solange die Frage ungeklärt ist oder geprüft wird, 2020–22
× Lehnt eine eigenständige gewohnheitsrechtliche Regel ab	
 Vereinigte Staaten	"Hat die Staatenpraxis und die opinio juris nicht festgestellt" ⁵⁴ für eine allgemeine Verpflichtung (2021), bekräftigt in der Debatte des Sechsten Ausschusses 2025 zum neuen Thema der ILC zur Sorgfaltspflicht – akzeptiert jedoch, dass ein notifizierter Staat angemessene Schritte ergreifen muss; Lex-specialis-Argumentation im Klimaverfahren
 Vereinigtes Königreich	"Noch keine ausreichende Staatenpraxis, um eine spezifische Regel des Völkerge- wohnheitsrechts zu begründen" ⁵⁵ , 2021
 Israel ·  Argentinien	Nur freiwillige, unverbindliche Norm; keine opinio juris, 2020–21

Tabelle 2. Positionen der Staaten zur Sorgfaltspflicht als verbindliche gewohnheitsrechtliche Verpflichtung, wie sie in nationalen und gemeinsamen Positionen zum Völkerrecht im Cyberraum (2019–2024), im Klimagutachtenverfahren (2024–25) und in der Debatte des Sechsten Ausschusses zum Thema der ILC "Due diligence in international law" (2025) zum Ausdruck kommen.

Wichtig ist, dass die Meinungsverschiedenheit über die Einordnung der Sorgfaltspflicht geringer ist, als es scheinen mag. Die USA lehnen die allgemeine Verpflichtung vor allem aus der Befürchtung ab, Gerichte könnten eine allgemeine Verpflichtung zu extensiv in neue Bereiche hineinlesen. Doch selbst skeptische Staaten wie die USA akzeptieren regelmässig das inhaltliche Verhalten, das die Sorgfaltspflicht verlangt, sobald es für einen neuen Bereich näher konkretisiert worden ist. So lehnen die USA zwar das eigenständige Prinzip ab, akzeptieren aber ausdrücklich, dass "Staaten, die Tätigkeiten im Cyberraum ausüben, die Souveränität anderer Staaten berücksichtigen müssen, auch ausserhalb des Kontexts bewaffneter Konflikte"⁵⁶, und dass "ein Staat, der über eine von seinem Territorium ausgehende schädliche Tätigkeit benachrichtigt wird, angemessene Schritte ergreifen muss, um gegen diese Tätigkeit vorzugehen".⁵⁷ Die USA haben zudem selbst mehrere der grundlegenden Präzedenzfälle der Doktrin als erfolgreiche Klägerin geschaffen – nicht nur im Fall der Alabama Claims, sondern auch im Trail-Smelter-Fall gegen Kanada.⁵⁸ Ein Anwendungsprozess könnte somit dennoch konkretisieren, was die Sorgfaltspflicht beim Kontrollverlust verlangt, und zwar in einer Form, die ein Staat für diesen spezifischen Bereich mittragen kann, ohne das eigenständige Prinzip anzuerkennen.

2.3 Elemente der Sorgfaltspflicht

Die Sorgfaltspflicht ist nicht optional, aber ihre Anforderungen variieren je nach Bereich. Zu den Elementen, die sich je nach Kontext unterscheiden, gehören:

⁵⁴ Englisch Original: "Has not identified the State practice and opinio juris".

⁵⁵ Englisch Original: "Not yet State practice sufficient to establish a specific customary international law rule".

⁵⁶ Koh, H. H. (2012, September 18). *International law in cyberspace [Speech transcript]*. USCYBERCOM Inter-Agency Legal Conference, Fort Meade, MD. Opinio Juris. Englisch Original: "States conducting activities in cyberspace must take into account the sovereignty of other States, including outside the context of armed conflict".

⁵⁷ 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. Englisch Original: "if a State is notified of harmful activity emanating from its territory it must take reasonable steps to address such activity".

⁵⁸ *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.

Kenntnis als Auslöser: Die Sorgfaltspflicht wird durch tatsächliche Kenntnis oder Kennenmüssen eines Risikos ausgelöst. Wie im Fall Bosnien gegen Serbien festgehalten: Die Handlungspflicht entsteht "in dem Augenblick, in dem der Staat vom Bestehen einer ernsthaften Gefahr erfährt oder normalerweise hätte erfahren müssen."⁵⁹ Für das Vorliegen eines Risikos ist keine vollständige wissenschaftliche Gewissheit erforderlich. Wie das Klimagutachten des IGH ergänzt, genügt es nicht, dass sich Staaten pauschal auf wissenschaftliche Unsicherheit berufen, um Sorgfaltspflichten zu bestreiten.⁶⁰

Mindestschwelle für die Auslösung: Die Sorgfaltspflicht soll risikoproportional sein. Sie wird nicht bei jeder geringfügigen Unannehmlichkeit ausgelöst. Die Kodifizierung im Umweltbereich legt die Mindestschwelle für die Auslösung der Sorgfaltspflicht bei "erheblichen" Schäden fest.⁶¹ Sie definiert die erfassten Risiken so, dass sie sowohl "eine hohe Wahrscheinlichkeit, erhebliche grenzüberschreitende Schäden zu verursachen, als auch eine geringe Wahrscheinlichkeit, katastrophale grenzüberschreitende Schäden zu verursachen"⁶² umfassen.⁶³ Schäden umfassen Schäden an Personen, Sachen oder der Umwelt an Orten unter der Hoheitsgewalt eines anderen Staates.⁶⁴

Strenge: Der Sorgfaltsmassstab "kann sich in Bezug auf die mit der Tätigkeit verbundenen Risiken ändern"⁶⁵ und kann sich mit zusätzlichem wissenschaftlichem und technologischem Wissen im Laufe der Zeit verschärfen oder lockern.⁶⁶ Bei ultragefährlichen Tätigkeiten steigt der Sorgfaltsmassstab und verlangt "ein viel höheres Mass an Nachdruck seitens des Staates bei deren Durchsetzung".⁶⁷ Der IGH stellte 2025 zudem fest, dass der Massstab für den Klimawandel als Risiko von "im Wesentlichen universeller" Tragweite "streng" ist und "ein erhöhtes Mass an Wachsamkeit und Prävention"⁶⁸ verlangt.⁶⁹

Präventionspflicht vs. Beendigungspflicht: Es kann eine Pflicht geben, Schaden zu verhindern, bevor er eintritt, und eine Pflicht, Schaden zu beenden, sobald ein Staat davon erfährt. Im Tallinn-Handbuch liegt die Last fast ausschliesslich auf der Beendigung bei Kenntniserlangung. Bei den höheren Stufen des Kontrollverlusts ist eine nachträgliche Beendigung hingegen sehr schwierig oder sogar unmöglich.

⁵⁹ *Bosnia v Serbia*, ICJ Reports 2007, para. 431. Englisch Original: "at the instant that the State learns of, or should normally have learned of, the existence of a serious risk."

⁶⁰ *Obligations of States in respect of Climate Change*, ICJ, Advisory Opinion, 23 July 2025 (nachfolgend *Climate AO*), paras. 293–294 und zugehörige Zusammenfassung.

⁶¹ ILC Prevention Articles (2001), commentary to Art. 2, para. (4).

⁶² Englisch Original: "a high probability of causing significant transboundary harm and a low probability of causing disastrous transboundary harm".

⁶³ ILC Prevention Articles (2001), Art. 2(a) and commentary: Eine ultragefährliche Tätigkeit beinhaltet "eine Gefahr, deren Eintritt selten zu erwarten ist, die aber in diesem seltenen Fall gravierende (mehr als erhebliche, ernsthafte oder beträchtliche) Ausmasse annehmen kann." (englisches Original: "a danger that is rarely expected to materialize but might assume, on that rare occasion, grave (more than significant, serious or substantial) proportions.")

⁶⁴ ILC Prevention Articles (2001), [Art. 2b & 2c](#)

⁶⁵ Englisch Original: "may change in relation to the risks involved in the activity".

⁶⁶ *Seabed Advisory Opinion* (2011), para. 117: Die Sorgfaltspflicht ist "ein variables Konzept", das sich "im Laufe der Zeit ändern kann" (englisches Original: "may change over time"), und zwar auch "im Verhältnis zu den mit der Tätigkeit verbundenen Risiken." (englisches Original: "in relation to the risks involved in the activity.")

⁶⁷ ILC Prevention Articles (2001), commentary to Art. 3, para. (11). Englisch Original: "much higher degree of vigour on the part of the State to enforce them".

⁶⁸ Englisch Original: "a heightened degree of vigilance and prevention".

⁶⁹ *Climate AO*, paras. 137–138.

Absolute Pflicht vs. verfügbare Mittel: Die ILC behandelt das "wirtschaftliche Niveau der Staaten"⁷⁰ als "einen der zu berücksichtigenden Faktoren"⁷¹ für die Sorgfaltspflicht, weist aber zugleich darauf hin, dass das wirtschaftliche Niveau eines Staates "nicht dazu herangezogen werden kann, den Staat von seiner Verpflichtung zu entbinden"⁷². Die Meeresbodenkammer verlangt gleiche Standards für entwickelte Staaten und Entwicklungsstaaten, um "befürwortende Gefälligkeitsstaaten"⁷³ (sponsoring states of convenience) zu verhindern.⁷⁴ Das Klimagutachten des IGH legt dar, dass Kapazitäten ein Faktor für den anwendbaren Massstab sind, jedoch nur oberhalb einer nicht verhandelbaren Untergrenze.

Private Infrastruktur: Die Pflicht knüpft grundsätzlich an das Territorium eines Staates und an Tätigkeiten unter seiner "Hoheitsgewalt oder Kontrolle" an, einschliesslich rechtmässigen, gesellschaftlich wertvollen Verhaltens, das Risiken birgt.⁷⁵ Wie weit die Pflicht in private Infrastruktur hineinreicht und was sie von Staaten verlangt, durch die ein entwichenes System lediglich hindurchgeleitet wird, müsste noch bestimmt werden.⁷⁶ Grundsätzlich bedürfen gefährliche Tätigkeiten innerhalb einer Hoheitsgewalt jedoch der vorherigen Genehmigung durch den Staat, und jede Genehmigungsentscheidung "stützt sich ... auf eine Bewertung der möglichen grenzüberschreitenden Schäden."⁷⁷ Ferner versteht die ILC die Präventionspflicht als fortlaufenden Prozess und nicht als einmalige Beurteilung. Sie kann die Notifikation von Risiken, Konsultationen über Präventionsmassnahmen, einen fortlaufenden Informationsaustausch und Notfallnotifikationen umfassen.⁷⁸

2.4 Der breitere Governance-Kontext der Sorgfaltspflicht

Die internationalen Regeln für KI werden nicht an einem einzigen Ort festgelegt. Ein Bericht des Eidgenössischen Departements für auswärtige Angelegenheiten (EDA) aus dem Jahr 2022 kartierte den entstehenden Rahmen über fünf Ebenen und kam zum Schluss, dass die entscheidende Variable nicht die Aktivität auf einer einzelnen Ebene ist, sondern die Stärke der Verbindungen zwischen ihnen.⁷⁹ Die fünf Ebenen sind die folgenden:

1. **Völkerrecht:** Verträge und gewohnheitsrechtliche Regeln, die Staaten binden. Hier ist die in diesem Bericht analysierte Sorgfaltspflicht angesiedelt.

⁷⁰ Englisch Original: "economic level of States".

⁷¹ Englisch Original: "one of the factors to be taken into account".

⁷² ILC Prevention Articles (2001), commentary to Art. 3, para. (13). Englisch Original: "cannot be used to dispense the State from its obligation".

⁷³ Englisch Original: "sponsoring states of convenience".

⁷⁴ Seabed Advisory Opinion (2011), paras. 158–159.

⁷⁵ ILC Prevention Articles (2001), [Art. 1](#)

⁷⁶ ILC Prevention Articles (2001), Art. 1 (Tätigkeiten "im Hoheitsgebiet oder anderweitig unter der Hoheitsgewalt oder Kontrolle" (englisches Original: "in the territory or otherwise under the jurisdiction or control") eines Staates); Tallinn Manual 2.0 (2017), Rule 6.

⁷⁷ ILC Prevention Articles (2001), Arts. 6–7. Englisch Original: "shall ... be based on an assessment of the possible transboundary harm."

⁷⁸ ILC Prevention Articles (2001), Arts. 4, 8, 9, 12, 17

⁷⁹ Federal Department of Foreign Affairs. (2022) "Artificial intelligence and international rules", Report to the Federal Council, Bern.

2. **Soft Law:** Die KI-Grundsätze der OECD, die UNESCO-Empfehlung zur Ethik der KI, Erklärungen der G7 und von Gipfeltreffen – Instrumente mit normativem Gewicht, aber ohne bindende Kraft.
3. **Nationales Recht mit De-facto-internationaler Reichweite:** Dazu gehört vor allem der AI Act der EU, aber auch nationale Regelwerke, deren Marktmacht ihre Anforderungen exportiert.
4. **Internationale Standards:** Dazu gehören Gremien mit einem ganzen Spektrum staatlicher Beteiligung. Die ITU ist eine Sonderorganisation der UNO, deren Mitglieder die Staaten selbst sind. Die Mitglieder von ISO und IEC sind nationale Normungsorganisationen, deren Zusammensetzung von Land zu Land variiert. Daneben gibt es private Normungsgremien mit Einzelmitgliedern wie die IETF und Konsortien mit Unternehmensmitgliedern wie die Linux Foundation.
5. **Unternehmenspraxis:** Die Frontier-KI-Sicherheitsrahmenwerke, Evaluationsregime und internen Richtlinien, an die sich KI-Entwickler selbst binden.

Ein Beispiel dafür, wie die Sorgfaltspflicht mit den Ebenen 2–5 verknüpft ist

Das allgemeine Rechtsprinzip der Sorgfaltspflicht, wie es an anderer Stelle in diesem Bericht verwendet wird, ist stets auf Ebene 1 angesiedelt und gilt stets für Staaten über ihr Territorium. Die Sorgfaltspflicht kann jedoch mit anderen Ebenen interagieren, und mitunter entlehnen diese Ebenen den Begriff "Due Diligence" sogar aus dem Völkerrecht, wie dies bei der "menschenrechtlichen Sorgfaltspflicht"⁸⁰ (Human Rights Due Diligence) der Fall ist. Konkret wurde im Kontext von Wirtschaft und Menschenrechten der zwischenstaatliche Begriff "Due Diligence" auch zur Beschreibung unternehmerischer Risikomanagementprozesse verwendet. Terminologisch ist das etwas verwirrend, aber es ist beabsichtigt: Freiwillige Leitlinien (Ebene 2) und nationales Recht mit internationaler Reichweite (Ebene 3), die für Unternehmen gelten, sollen so als Operationalisierungen übergeordneter Pflichten aus dem Völkerrecht verstanden werden.

Die UNO-Leitprinzipien für Wirtschaft und Menschenrechte halten fest, dass Staaten nach dem Völkerrecht grundsätzlich Schutz vor Menschenrechtsverletzungen "innerhalb ihres Hoheitsgebiets und/oder ihrer Jurisdiktion durch Dritte, einschliesslich Wirtschaftsunternehmen"⁸¹ gewähren müssen (Ebene 1, Prinzip 1). Darauf aufbauend bieten die Leitprinzipien Soft-Law-Leitlinien, wonach Staaten, die Wirtschaftsunternehmen besitzen, kontrollieren oder wesentlich unterstützen, "zusätzliche Schritte unternehmen sollten"⁸², um vor Verletzungen zu schützen, "gegebenenfalls auch, indem sie menschenrechtliche Sorgfaltspflicht verlangen"⁸³ (Ebene 2, Prinzip 4) – was in diesem Fall bedeutet, Unternehmen in die Verantwortung zu nehmen, menschenrechtliche Auswirkungen zu ermitteln, zu verhüten und zu mindern (Ebene 2, Prinzipien 15–17).⁸⁴ Einige Rechtsordnungen haben solche Leitlinien durch nationales Recht mit internationaler Reichweite verbindlich gemacht (Ebene

⁸⁰ Englisch Original: "human rights due diligence".

⁸¹ Englisch Original: "within their territory and/or jurisdiction by third parties, including business enterprises".

⁸² Englisch Original: "should take additional steps".

⁸³ Englisch Original: "including, where appropriate, by requiring human rights due diligence".

⁸⁴ 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.

3), und das B-Tech-Projekt der UNO hat mit KI-spezifischen Leitlinien zur Anwendung der menschenrechtlichen Sorgfaltspflicht auf generative KI nachgezogen (Ebene 2).⁸⁵

Entscheidend ist hier nicht der menschenrechtsspezifische Inhalt, sondern dass ein Zusammenspiel der Ebenen besteht. Eine Sorgfaltspflicht aus dem Völkerrecht gilt für Staaten. Operativ kann diese Pflicht jedoch erfüllt werden, indem von privaten Akteuren ein unternehmerischer Risikomanagementprozess verlangt wird, wobei die konkreten Verfahren durch Soft Law und Standards bereitgestellt werden und die Durchsetzung mit der Zeit möglicherweise ins nationale Recht übergeht.

Die Verbindungen können auch nach oben verlaufen

Das Menschenrechtsbeispiel verläuft vom Völkerrecht nach unten. Verbindungen können aber auch von der Praxis nach oben zum Recht verlaufen. ISO/IEC 38507:2022 gibt Leitungsgremien Orientierung für die Governance der KI-Nutzung ihrer Organisationen. Die Norm macht Unternehmenspraxis (Ebene 5) zu einem internationalen Standard (Ebene 4). Ihr Kernprinzip lautet, dass das Leitungsorgan "für alle Tätigkeiten einer Organisation rechenschaftspflichtig bleibt. Diese Rechenschaftspflicht kann nicht delegiert werden"⁸⁶, und dass Mitglieder des Leitungsgremiums die Verantwortung für die KI-Nutzung übernehmen sollten, "anstatt die Verantwortung dem KI-System selbst zuzuschreiben".⁸⁷

Solche Standards können nach oben wandern. Regulierungsbehörden und Gerichte (Ebene 3) sowie Soft Law (Ebene 2) ziehen sie als Massstab für angemessene Aufsicht heran. Auf Ebene 1 ergeben sich Sorgfaltspflichten aus völkerrechtlichen Quellen, sie können sich jedoch mit wissenschaftlichem Wissen verschärfen, und eine weithin übernommene Praxis kann zum Beleg dafür werden, was die Sorgfaltspflicht verlangt. So stellte der IGH im Fall *Pulp Mills* fest, dass die Umweltverträglichkeitsprüfung zu einem Erfordernis des allgemeinen Völkerrechts geworden war, weil sie unter den Staaten breite Akzeptanz gefunden hatte.⁸⁸ Für den Kontrollverlust stärkt die weithin akzeptierte Tatsache, dass das Leitungsgremium eines Unternehmens die Rechenschaftspflicht nicht auf ein KI-System abwälzen kann, das Argument, dass sich auch ein Staat nicht auf die Autonomie eines Systems berufen kann, um sich seiner Präventionspflicht bei erheblichen grenzüberschreitenden Risiken zu entziehen.

Ein dualer Prozess aus Anwendung und Operationalisierung

Regierungen spielen nicht auf allen Ebenen eine gleich zentrale Rolle. Die Ebenen eins und zwei werden weitgehend durch zwischenstaatliche Verhandlungen und die Auslegung durch Rechtsfachleute geprägt. Auf Ebene drei handelt eine mächtige Regierung im Alleingang. Auf Ebene vier sind Regierungen nur indirekt präsent, und nur, wenn sie sich dafür entscheiden. Auf Ebene fünf hat der Staat keinen formellen Sitz, auch wenn freiwilliges privates Verhalten dennoch von Erwartungen an künftige Regulierung, Beschaffung oder Haftung geprägt sein kann.

⁸⁵ 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.

⁸⁶ Englisches Original: "remains accountable for all activities of an organization. This accountability cannot be delegated".

⁸⁷ ISO/IEC 38507:2022, *Governance implications of the use of artificial intelligence by organizations*, ISO/IEC JTC 1/SC 42. Englisches Original: "rather than attributing responsibility to the AI system itself".

⁸⁸ *Pulp Mills (Argentina v. Uruguay)*, ICJ Reports 2010, para. 204.

Die Präventionspflicht ist ein Instrument der Ebene eins. Die übergeordnete Orientierung darüber, **wie die rechtlichen Elemente der Sorgfaltspflicht**, etwa Umfang und Strenge, auf einen Bereich anzuwenden sind, ist eine Frage der Ebene 1, die unter Völkerrechtsfachleuten zu klären ist.

Die **operative Umsetzung – also welche konkreten Verfahren erforderlich sind, um die rechtliche Pflicht zu erfüllen**, einen Kontrollverlust zu verhindern – profitiert hingegen vom Einbezug der Ebenen 2–5 und damit auch von einer stärkeren Beteiligung des Privatsektors. So stammte dieser operative Inhalt im Umweltbereich aus Umweltverträglichkeitsprüfungen, bei der Bekämpfung der Terrorismusfinanzierung aus der Evaluationsmethodik der Financial Action Task Force (FATF).

Anwendung und Operationalisierung der Sorgfaltspflicht können entweder nacheinander oder parallel erfolgen, mit sich überschneidenden, aber nicht deckungsgleichen Kreisen von Stakeholdern für beide.

3. Bereichsübergreifender Vergleich

In diesem Abschnitt vergleichen wir die Anwendung der Sorgfaltspflicht in vier verschiedenen Bereichen: Umwelt, Cyber, Terrorismus und Menschenrechte.

	Umwelt	Cyber	Terrorismus / NSA	Menschenrechte
Rechtsgrundlage	Gewohnheitsrecht (IGH, einstimmig); UNCLOS; MEA	Umstrittenes Gewohnheitsrecht; freiwillige UNO-Norm 13(c)	Kap. VII (Res. 1373); Gewohnheitsrecht; Verträge	Menschenrechtsverträge; Völkermordkonvention Art. I
Auslöser	Risiko erheblicher Schäden, einschl. kumulativer sowie unwahrscheinlicher, katastrophaler Risiken	Tatsächliche Kenntnis oder Kennenmüssen; keine aus der Souveränität abgeleitete Vermutung der Kenntnis	Kenntnis + Kapazität (Teheran); Unvermögen entlastet (DRK gegen Uganda)	„Wusste oder hätte wissen müssen“ ⁸⁹ von einer realen und unmittelbaren Gefahr; ernsthafte Gefahr (Völkermord)
Massstab	„Streng“; erhöhte Wachsamkeit; objektiv, nach Kapazität abgestuft	Verhalten; zumutbare Massnahmen; keine Überwachungspflicht	Verhalten; angemessene Massnahmen im Rahmen der Kapazität	Angemessene Massnahmen; keine unmögliche oder unverhältnismässige Last; qualitätsgeprüft
Prozedurales Rückgrat	UVP, Notifikation, Konsultation, Zusammenarbeit (gewohnheitsrechtlich)	Dünn: Erwartungen zu Notifikation/Unterstützung (GGE 2021)	Berichterstattung an CTC/CTED; FATF-Evaluationen	Risikobewertung, Rahmen- und operative Pflichten, Untersuchung
Klagebefugnis	Erga omnes (Klimagutachten ¶440)	Nur verletzter Staat	UNO-Sicherheitsrat; verletzter Staat	Individualbeschwerde; erga omnes partes (Völkermord)
Durchsetzungsbilanz	Feststellungsurteile; Weg zur Wiedergutmachung 2025 eröffnet; intensive Aufnahme in Rechtsstreitigkeiten	Keine in concreto; Gegenmassnahmen theoretisch erörtert	Sanktionen, Listung, Gewalt; kaum gerichtliche Feststellungen	Regelmässige Feststellung von Verletzungen; bescheidene Abhilfen

Tabelle 3. Vergleichsmatrix. „NSA“ = nichtstaatliche bewaffnete Akteure.

⁸⁹ Englisch: Original: „Knew or ought to have known“.

Bewertet man dasselbe Material auf einer Fünf-Punkte-Skala (5 = am stärksten), werden die Profile auf einen Blick lesbar:

Dimension der Stärke	Umwelt	Cyber	Terrorismus	Menschenrechte
Gewohnheitsrechtliche Verbindlichkeit (gefestigt?)	5	2	4	5
Bestimmtheit des Standards	4	3	3	5
Prozeduraler Gehalt (Prüfung/Notifikation)	5	2	3	4
Gerichtliche Durchsetzbarkeit	3	1	1	5
Praktische Befolgungswirkung	3	2	4	4

Tabelle 4. Stärke der Präventionspflicht nach Bereich, Einschätzung des AGIPI, Juli 2026. Die Bewertungen beruhen auf qualitativer Beurteilung, nicht auf einer formalen Metrik.

3.1 Was erklärt die Unterschiede?

Die Bestimmung dessen, was die Sorgfaltspflicht verlangt, "erfordert eine Beurteilung in concreto"⁹⁰ dessen, was unter den spezifischen Umständen des jeweiligen Bereichs letztlich angemessen ist.⁹¹ Zu den etablierten Faktoren für die Beurteilung des Bestehens und der Strenge einer Sorgfaltspflicht in einem neuen Kontext gehören:⁹²

- wissenschaftliche und technologische Erkenntnisse
- internationale Regeln und Standards
- das Schadensrisiko
- die Dringlichkeit
- die verfügbaren Mittel und Fähigkeiten des betreffenden Staates

Allerdings lassen sich diese Unterschiede nicht allein durch Risikomerkmale erklären. Auch Terrorismus birgt Risiken mit hoher Schadenswirkung, und doch gab es kaum Rechtsstreitigkeiten wegen unzureichender Terrorismusprävention. Offensichtlich erfahren manche Bereiche weit mehr rechtliche Durchsetzung als andere.

Institutionalisiertes Risikobewusstsein: Die Sorgfaltspflicht wird durch Wissen ausgelöst. Das Klimarecht hat sich unter anderem deshalb verfestigt, weil der IPCC ein Bestreiten der Vorhersehbarkeit unhaltbar gemacht hat: Der IGH stützte sich direkt auf den wissenschaftlichen Konsens, um festzuhalten, dass Staaten die einschlägige Wissenschaft "aktiv verfolgen" müssen und sich nicht hinter Unsicherheit verstecken können. Die Terrorismusbekämpfung beruht auf nachrichtendienstlichen Erkenntnissen statt auf geteilter Wissenschaft, wodurch das Wissen asymmetrisch, bestreitbar und schwer justiziabel ist.

⁹⁰ Englisch Original: "calls for an assessment in concreto".

⁹¹ Application of the Genocide Convention (Bosnia and Herzegovina v Serbia and Montenegro), ICJ Reports 2007, para. 430 (die Sorgfaltspflicht "erfordert eine Beurteilung in concreto" (englisches Original: "calls for an assessment in concreto")).

⁹² Obligations of States in respect of Climate Change, ICJ, Advisory Opinion, 23 July 2025, paras. 282–300

Wissen im Cyberbereich wird bewusst nicht geteilt. Ein Faktor, der die Präventionspflicht stärkt, ist daher eine gemeinsame, massgebliche, von Staaten getragene Risikobewertung.

Die Position der fähigsten Staaten: Wo die Staaten mit der grössten Fähigkeit, den Schaden zu verursachen, ihm zugleich in hohem Masse ausgesetzt sind, akzeptieren sie die Pflicht (Umwelt – jede grosse Volkswirtschaft hat eine Küste und ein Klima). Wo die fähigen Staaten in erster Linie Standort der schädlichen Fähigkeit sind und nicht deren Opfer, leisten sie Widerstand (Cyber: Die USA, das Vereinigte Königreich und Israel sind die beharrlichen Verweigerer – “persistent objectors”). Eine Präventionspflicht kristallisiert sich heraus, wenn die Vetospieler zum Schluss kommen, dass auch sie Opfer sind.

Beobachtbarkeit von Verletzungen: Umweltschäden sind messbar und zunehmend zurechenbar. Cyber-Fahrlässigkeit ist von aussen unsichtbar, und Terrorismusfälle würden von umstrittenen nachrichtendienstlichen Erkenntnissen abhängen. Je eher sich Sorgfaltsverstösse in einem Bereich durch öffentliche, objektive Beweise feststellen lassen, desto stärker erfolgt die Durchsetzung auf gerichtlichem Weg. Ohne Beobachtbarkeit verlagert sich die Durchsetzung stärker auf Wirtschaftssanktionen, Technologiesanktionen und Gewaltanwendung.

Wer klagebefugt ist: Die Menschenrechte sind das am stärksten durchgesetzte Rechtsgebiet, weil Millionen von Einzelpersonen klagen können. Das Umweltrecht erhielt 2025 eine Klagebefugnis erga omnes, wodurch sich die Zahl potenzieller Anspruchsteller vervielfacht. In den Bereichen Cyber und Terrorismus bleibt die Beschwerde auf geschädigte Staaten beschränkt, die Rechtsansprüche gegen diplomatische Kosten abwägen und meist schweigen.

Von Prinzipien zu Verfahren: Pflichten, die auf Ebene 1 in allgemeinen Begriffen formuliert sind (“geeignete Massnahmen ergreifen”), sind nach wie vor schwer justiziabel. Pflichten, die auf den Ebenen 2–5 in einzelne, datierte Handlungen prozeduralisiert werden, erzeugen nachweisbare Verletzungen und bürokratische Routinen. Der Apparat der Umweltverträglichkeitsprüfung (UVP) und der Evaluationszyklus der FATF zeigen: Konkrete Verfahren, die Einhaltung und Nichteinhaltung sichtbar machen, verwandeln die Sorgfaltspflicht von einer abstrakten Verpflichtung in eine nachprüfbare Aktenlage.

Zusammenfassend erreicht die Präventionspflicht ihr Maximum, wo ein massgebliches epistemisches Gremium das Wissen um das Risiko universalisiert hat, die fähigen Staaten sich selbst zu den Exponierten zählen, Verletzungen öffentlich beobachtbar sind, die Klagebefugnis breit ist und die Pflicht prozeduralisiert wurde. Ihr Minimum erreicht sie, wo das Wissen umstritten oder geheim ist, die fähigen Staaten die Gefahrenquelle beherbergen, Verletzungen unsichtbar sind, nur Staaten klagen können und die Pflicht unzureichend operationalisiert bleibt. Im Umweltrecht ist die Pflicht am stärksten, im Cyberbereich am schwächsten.

4. Anwendung der Sorgfaltspflicht auf den Kontrollverlust

Dieser Abschnitt untersucht, wie die Sorgfaltspflicht auf den Kontext des Kontrollverlusts angewendet werden könnte. Er ersetzt nicht den eigentlichen Prozess, in dem Rechtsfachleute diese Arbeit leisten, kann aber einen grundlegenden Überblick bieten.

4.1 Vorläufige Beurteilung der Elemente der Sorgfaltspflicht

Das Kontrollverlustrisiko erreicht die Mindestschwelle für die Aktivierung: Die Risikodefinition der ILC umfasst ausdrücklich "eine geringe Wahrscheinlichkeit, katastrophalen grenzüberschreitenden Schaden zu verursachen".⁹³ Die höheren Stufen des Kontrollverlustrisikos entsprechen dieser Beschreibung. Viele Regierungen verfügen über nationale Risikoanalysen, die vorhersehbare gesellschaftliche Risiken abdecken, doch die dafür verwendeten Standardmethoden stossen bei Risiken mit extremer Schadenswirkung und bei neu entstehenden Risiken an strukturelle Grenzen,⁹⁴ und genau dort ist das Risiko des KI-Kontrollverlusts angesiedelt. Dass der Kontrollverlust in einer nationalen Risikoanalyse fehlt, ist daher kein Beleg dafür, dass ein Staat das Risiko geprüft und als unterhalb der Schwelle liegend befunden hat.

Das Kontrollverlustrisiko erreicht die Schwelle für erhöhte Strenge: Dasselbe Risikoprofil berührt den Zweig der Doktrin für ultragefährliche Tätigkeiten, der daran "einen wesentlich höheren Sorgfaltsmassstab bei der Ausgestaltung von Politiken und ein wesentlich höheres Mass an Nachdruck ... bei deren Durchsetzung"⁹⁵ knüpft⁹⁶. Wie das Alabama-Schiedsgericht ausführte, ist Sorgfalt "in genauem Verhältnis zu den Risiken"⁹⁷ geschuldet. Daraus folgt, dass die höheren Stufen des Kontrollverlusts am strengen Ende der Skala zu verorten sind.

Präventions- vs. Beendigungspflicht: Die Tallinn-Regel ist streng genommen eine Pflicht zur *Beendigung*: Ein Staat, der von einer schädigenden Operation Kenntnis erlangt, muss "alle unter den gegebenen Umständen möglichen Massnahmen ergreifen, um ihr ein Ende zu setzen".⁹⁸ Für behebbare Vorfälle ist die Beendigung ab Kenntnis eine tragfähige Untergrenze. Für einen irreversiblen Kontrollverlust ist sie strukturell unzureichend. Wir können nicht erst handeln, wenn sich ein hinreichend fähiges System exfiltriert, repliziert oder festgesetzt hat, da es dann sehr kostspielig oder unmöglich sein kann, es zu stoppen. Die rechtliche Konsequenz lautet, dass bei dieser Gefahr das Gewicht der Pflicht auf der Prävention liegen muss.

⁹³ ILC Articles on Prevention of Transboundary Harm (2001), Art. 2(a). Englisch Original: "a low probability of causing disastrous transboundary harm".

⁹⁴ Kohler, K. (2023). *National risk assessments of cross-border risks* (Risk and Resilience Report). Center for Security Studies, ETH Zurich.

⁹⁵ Englisch Original: "a much higher standard of care in designing policies and a much higher degree of vigour ... to enforce them".

⁹⁶ ILC Prevention Articles (2001), Art. 2(a); commentary to Art. 3, para. (11).

⁹⁷ Englisch Original: "in exact proportion to the risks".

⁹⁸ Tallinn Manual 2.0 (2017), Rule 7 and commentary to Rules 6–7; UN GGE Report 2021, A/76/135, paras. 29–30. Englisch Original: "take all measures that are feasible in the circumstances to put an end to".

Allgemeine Überwachung vs. gezielte Aufsicht: Der Konsens im Cyberbereich lehnt eine allgemeine Überwachungspflicht ab. Den gesamten IKT-Datenverkehr innerhalb eines Hoheitsgebiets zu überwachen, ist nahezu unmöglich und käme einem Freipass für einen Überwachungsstaat gleich.⁹⁹ Keine dieser beiden Begründungen lässt sich jedoch auf den Kontrollverlust übertragen. Die relevante Tätigkeit wird von einer Handvoll identifizierbarer Betreiber in registrierten Anlagen ausgeübt, die feststellbare Mengen an Rechenleistung verbrauchen. Zu überwachen, ob diese die richtigen Prozesse eingerichtet haben, ist weder unmöglich noch ein Problem für die bürgerlichen Freiheiten. Es gleicht eher der Aufsicht über gefährliche Industrien, etwa bei nuklearer Sicherheit und Biosicherheit.

Absolute Pflicht vs. "verfügbare Mittel": Für die höchste Stufe des Kontrollverlustrisikos sollte die Pflicht am absoluten Ende der Skala angesiedelt sein, doch die "verfügbaren Mittel" können relevanter sein, wenn es darum geht, die Auswirkungen von Vorfällen der Stufen L2 bis L3 zu mindern, bei denen KI-Modelle Infrastruktur in Drittstaaten nutzen.

Staatlich kontrollierte vs. private Infrastruktur: Seit den britischen Werften im Jahr 1872 geht es bei der Pflicht um die Regulierung privaten Verhaltens, und 2025 bekräftigte der IGH, dass ein Staat verantwortlich wird, wenn er gegenüber privaten Akteuren unter seiner Hoheitsgewalt "keine gebührende Sorgfalt walten liess, indem er die erforderlichen regulatorischen und gesetzgeberischen Massnahmen nicht ergriff".¹⁰⁰ In ähnlicher Weise verlangt *Pulp Mills* Wachsamkeit und administrative Kontrolle gegenüber "öffentlichen und privaten Betreibern"¹⁰¹ gleichermassen.¹⁰² Die Frage ist, wie weit die operativen Pflichten in private Infrastruktur hineinreichen.

Die Tallinn-Formulierung erfasst das Hoheitsgebiet eines Staates sowie "Cyberinfrastruktur unter seiner staatlichen Kontrolle"¹⁰³, doch Frontier-Rechenkapazität befindet sich ganz überwiegend in privater Hand. Zumindest die Untergrenze sollte unumstritten sein, da jeder Staat ein hoheitliches Eigeninteresse daran hat: **Jeder Staat ist mindestens verpflichtet zu verhindern, dass ein entwichenes oder unkontrolliertes KI-System seine nicht öffentlich zugänglichen, militärisch kontrollierten Systeme erreicht** – nukleare Führungs- und Kontrollsysteme, Waffensysteme, klassifizierte Netzwerke.

Tatsächliche Kenntnis oder Kennenmüssen? Die Positionen im Cyberbereich teilen sich in die Beschränkung auf tatsächliche Kenntnis und das Kennenmüssen (constructive knowledge), wobei allgemein Einigkeit besteht, dass Kenntnis nie allein aufgrund der Souveränität vermutet wird.¹⁰⁴ Für KI ist die Position des Kennenmüssens sinnvoller. Auf der Ebene der *Risikoklasse* haben zwei staatlich getragene internationale Expertengremien zu KI das Kontrollverlustrisiko aktenkundig gemacht: der International AI Safety Report, unterstützt von 30 Ländern, und das KI-Wissenschaftsgremium der UNO, das von über 190 Ländern in

⁹⁹ UN GGE Report 2021, A/76/135, para. 30(a); Österreichs Position von 2024 ergänzt eine eng gefasste Pflicht, "eine minimale staatliche IKT-Infrastruktur" (englisches Original: "a minimum governmental ICT infrastructure") zu unterhalten, die zur Erkennung fähig ist.

¹⁰⁰ *Climate AO*, para. 428. Englisches Original: "failed to exercise due diligence by not taking the necessary regulatory and legislative measures".

¹⁰¹ Englisches Original: "public and private operators".

¹⁰² *Pulp Mills*, ICJ Reports 2010, para. 197

¹⁰³ Englisches Original: "cyber infrastructure under its governmental control".

¹⁰⁴ *Corfu Channel*, ICJ Reports 1949, at 18; AU Common Position (2024), para. 23; nationale Positionen zusammengestellt im CCDCOE Cyber Law Toolkit, "Due diligence."

der UNO-Generalversammlung mandatiert wurde. Wie im Klimarecht kann sich ein Staat nicht hinter Unsicherheit verstecken, wenn eine massgebliche wissenschaftliche Bewertung, die er selbst mitmandatiert hat, das Risiko identifiziert hat.¹⁰⁵ Auf der Ebene des *konkreten Vorfalls* liegt das Wissen in der Regel beim Betreiber, nicht beim Staat. Wie das Wissenschaftsgremium der UNO festhält: „Relevante Beweise waren über OpenAI, Hugging Face und die von den Agenten genutzten Systeme verteilt.“¹⁰⁶ Der logische Lösungsweg ist eine Pflicht zur Meldung von Vorfällen und Fähigkeiten im Zusammenhang mit Kontrollverlust. Die vorfallspezifische Sorgfaltspflicht beginnt, sobald die Meldung fällig gewesen wäre, und nicht erst, wenn der Staat zufällig davon in den Nachrichten liest.

4.2 Grenzen

Noch keine KI-spezifische Staatenpraxis oder opinio juris. Die Rechtswissenschaft hat die Grundsätze aus dem Korfu-Kanal-Fall und das Schädigungsverbot (no-harm principle) auf KI-Technologien angewendet. Sie hat daraus Sorgfaltsmassnahmen abgeleitet und begonnen, die Analyse auf autonome KI-Agenten auszuweiten, einschliesslich Kontrollverlust und Selbsterhaltung als neu entstehende Risiken.¹⁰⁷ Bislang hat jedoch kein Staat eine nationale Position veröffentlicht, die ausdrücklich festhält, dass die Sorgfaltspflichtregel ihn in Bezug auf die KI-Entwicklung bindet. Die These lautet, dass die allgemeine gewohnheitsrechtliche Präventionspflicht auf KI ebenso anwendbar ist wie auf jede gefährliche Tätigkeit, wobei ihr Inhalt noch zu konkretisieren ist. Mit ausdrücklicher Unterstützung durch nationale Positionen wäre diese These stärker.

Die Schwäche der Sorgfaltspflichten bei von Menschen gesteuerten Cyberoperationen.

Die Staaten, in denen die Frontier-KI-Entwicklung stattfindet, sind weitgehend dieselben, die Sorgfaltspflichten im Cyberbereich blockiert haben. Generell neigen Standortstaaten der Entwicklung eher dazu, die Pflicht herunterzuspielen, während Opferstaaten sie wohl geltend machen würden. Das kann nur funktionieren, wenn sich die Frage hinreichend von der allgemeinen Debatte über Sorgfaltspflichten im Cyberbereich abgrenzen lässt – mit dem Argument, dass ein Kontrollverlust der Stufe L3 und insbesondere L4 grundlegend andere Risikomerkmale aufweist als das, was das Tallinn-Handbuch betrachtet hat, weil Kontrollverlust schwerwiegender und irreversibel sein kann und die Unterscheidung zwischen Standort- und Opferstaat aufhebt. Tatsächlich gilt: Wenn sich KI-Systeme

¹⁰⁵ *Climate AO*, para. 283; vgl. *Seabed Advisory Opinion* (2011), para. 131 (ein Staat „würde seine Sorgfaltspflicht nicht erfüllen, wenn er diese Risiken ausser Acht liesse“ (englisches Original: „would not meet its obligation of due diligence if it disregarded those risks“)).

¹⁰⁶ 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>. Englisches Original: „relevant evidence was spread across OpenAI, Hugging Face, and the systems used by the agents.“

¹⁰⁷ Coco & Dias (2024), at 240–243 (Anwendung der Grundsätze aus dem Korfu-Kanal-Fall und des Schädigungsverbots auf KI) und Abschnitt 12.5 (Massnahmen sorgfältigen staatlichen Verhaltens); Dias, T. (2025). *AI agents and global governance: Analyzing foundational legal, policy, and accountability tools*. Partnership on AI, pp. 8–11 und 16–18 (Sorgfaltspflicht bei grenzüberschreitenden Schäden durch KI-Agenten; Kontrollverlust und Selbsterhaltung unter den neu entstehenden Agentenrisiken; sowie die Frage der Autonomieschwelle, „jenseits derer Staaten die Entwicklung oder Bereitstellung von KI-Agenten nicht zulassen sollten“ (englisches Original: „beyond which states should not permit AI agents to be developed or deployed“)). Siehe auch 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> (mit dem Argument, dass Staaten bereits verpflichtet sind, globale Katastrophenrisiken zu mindern, einschliesslich solcher durch fortgeschrittene KI).

ausserhalb menschlicher Kontrolle ausbreiten und flächendeckend ein gesetzloseres Gleichgewicht in der Cybersicherheit schaffen, dann **sind digital fortgeschrittene Staaten infrastrukturell stärker exponiert**. Mehr Rechenzentren, die gehackt werden können, und eine stärkere Abhängigkeit von ihnen. Schliesslich verfügt KI, anders als der Cyberbereich, über die institutionalisierte, staatlich getragene Risikobewertung, die in Abschnitt 3.1 als Treiber einer stärkeren Durchsetzung identifiziert wurde.

4.3 Zusammenfassung

Das Recht, das grenzüberschreitende Kontrollverlustrisiken regelt, existiert bereits. Die **Sorgfaltspflicht (Due Diligence)** ist ein technologieneutraler Grundsatz des Völkerrechts, der an das Hoheitsgebiet anknüpft.

Das Risikoprofil des Kontrollverlusts verortet die Pflicht am strengen, ultragefährlichen Ende der Skala der Sorgfaltspflicht, und **bei potenziell irreversiblen Schaden muss ihr Gewicht auf der Prävention statt auf der Beendigung liegen**. Zwar ähnelt der Vorfall der Stufe L2 oberflächlich den Cyberoperationen, die das Tallinn-Handbuch beurteilt hat, doch das Kontrollverlustrisiko, einschliesslich L3 und L4, geht über diesen Rahmen hinaus. Die Cyberdebatte beurteilte behebbare Operationen nach einer Logik der Beendigung ab Kenntnis, nicht eine Gefahr, die ihre eigene Abhilfe vereiteln kann.

Mit dem International AI Safety Report und dem KI-Wissenschaftsgremium der UNO existiert ein Wissensauslöser für die Sorgfaltspflicht. Eine wissenschaftliche Bewertung, die von über 190 Ländern mandatiert ist und ein ernsthaftes Kontrollverlustrisiko feststellt, bedeutet, dass Regierungen vom Kontrollverlustrisiko tatsächlich Kenntnis haben oder haben müssten. Wie die Ko-Vorsitzende des KI-Wissenschaftsgremiums der UNO, Maria Ressa, erklärte: "Dieser Bericht ist der Moment, in dem die Welt nicht mehr sagen kann, sie habe es nicht gewusst."¹⁰⁸ Das Gremium fügt hinzu, dass der Kontrollverlust "genau die Art von Entscheidungsproblem darstellt, für die das Vorsorgeprinzip geschaffen wurde: eines, bei dem der potenzielle Schaden katastrophal oder irreversibel sein kann, auch wenn seine Wahrscheinlichkeit wissenschaftlich ungewiss bleibt."¹⁰⁹

Es besteht jedoch nach wie vor eine **Anwendungslücke**. Was fehlt, sind nationale Rechtspositionen, kodifizierte bewährte Praktiken und ein Prozess, der diese über geopolitische Bruchlinien hinweg zusammenführt, nach dem Vorbild des Montreux-Dokuments. Das AGI Preparedness Institute ist bereit, zur Schliessung dieser Lücke beizutragen. Hier in Genf, dem Geburtsort der Sorgfaltspflichtdoktrin.

¹⁰⁸ 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>.

Englisches Original: "This report is the moment the world stops being able to say it didn't know."

¹⁰⁹ 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>. Englisches Original: "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."

5. Operationalisierung der Sorgfaltspflicht bei Kontrollverlust

Dieser Abschnitt untersucht, wie die Sorgfaltspflicht bei Kontrollverlust unter Rückgriff auf die Ebenen 2–5 operationalisiert werden könnte. Er ersetzt nicht den eigentlichen Prozess, in dem Fachleute aus verschiedenen Anspruchsgruppen diese Arbeit leisten, kann aber einen grundlegenden Überblick bieten.¹¹⁰

5.1 Risikomanagementstandards als Orientierung

Die allgemeine Struktur des Risikomanagements ist in ISO 31000 festgelegt. Zu den KI-spezifischen Leitlinien für organisatorische Anforderungen gehören das Risikomanagement für KI (ISO/IEC 23894), ein zertifizierbares KI-Managementsystem (ISO/IEC 42001), die Folgenabschätzung (ISO/IEC 42005) und die Akkreditierung der Auditoren, welche die Zertifizierung glaubwürdig machen (ISO/IEC 42006).

Unterhalb dieser mittleren Ebene liefern operative Standards das, was eine Aufsichtsbehörde tatsächlich prüfen kann. ISO/IEC TS 8200 zur Steuerbarkeit automatisierter KI-Systeme behandelt die Beobachtbarkeit von Systemzuständen, Prozess und Kosten der Kontrollübergabe sowie die Reaktion auf Unsicherheit während der Kontrollübergabe. Daneben stehen die Datenelemente von Vorfallmeldungen (ISO/IEC AWI 25870, aufbauend auf dem gemeinsamen Meldeschema der OECD). Eine Sorgfaltspflicht bei Kontrollverlust muss ihre Compliance-Architektur nicht neu erfinden; sie muss auf diesen Stapel von Standards verweisen und die Schwellenwerte festlegen.

5.2 Ein vorläufiger Katalog spezifischer Massnahmen

- a) **Fähigkeitsbewertung vor der Bereitstellung anhand definierter Schwellenwerte.** Staaten sollten verlangen, dass Frontier-Modelle vor der Bereitstellung und an festgelegten Kontrollpunkten während des Trainings anhand kontrollverlustrelevanter Fähigkeitsschwellen wie Autonomie, Selbstreplikation oder Sabotage der Aufsicht evaluiert werden, wobei die Ergebnisse in einem Safety Case dokumentiert werden. Die Praxis existiert bereits: Die Frontier-Sicherheitsrahmenwerke von Anthropic, OpenAI und Google DeepMind definieren Fähigkeitsschwellen mit zugehörigen Risikominderungsmaßnahmen.¹¹¹ Externe Evaluationsstellen haben bereits im Rahmen von Vereinbarungen mit Entwicklern Bewertungen vor der Bereitstellung durchgeführt. In der Terminologie der ILC-Präventionsartikel entspricht dies den Artikeln 6 und 7: einer an eine Prüfung geknüpften Genehmigung.¹¹²
- b) **Kontrollmaßnahmen, die von Misalignment ausgehen.** Da eine Evaluation nicht bescheinigen kann, dass ein System korrekt ausgerichtet (aligned) ist, sollte die Bereitstellung fähiger Systeme unter Kontrollmaßnahmen erfolgen, die so ausgelegt

¹¹⁰ Für eine parallele Ableitung von Sorgfaltsmaßnahmen für KI direkt aus den Schutzpflichten – kontinuierliche Folgenabschätzungen über den gesamten Lebenszyklus der Technologie, wirksame menschliche Aufsicht, innerstaatliche Regulierung, Kapazitätsaufbau und internationale Zusammenarbeit – siehe Coco & Dias (2024), Abschnitt 12.5.

¹¹¹ Anthropic, Responsible Scaling Policy (aktuelle Fassung, 2025); OpenAI, Preparedness Framework, v2 (April 2025); Google DeepMind, Frontier Safety Framework, v3.0 (September 2025).

¹¹² ILC Prevention Articles (2001), Art. 6 (vorherige Genehmigung) und Art. 7 (Genehmigung auf Grundlage einer Prüfung möglicher grenzüberschreitender Schäden).

sind, dass sie selbst dann standhalten, wenn das Modell sie aktiv zu unterlaufen versucht. Die Forschungsgrundlage bildet die von Redwood Research entwickelte Control-Agenda¹¹³, und eine prüfbare Form existiert bereits im Control-Standard von Guidelight, dessen erste öffentliche Bewertung im August 2026 ergab, dass diese Praktiken bei den fünf führenden Entwicklern bestenfalls teilweise umgesetzt sind.¹¹⁴

- c) **Erhalt der Chain-of-Thought-Überwachbarkeit.** Bei heutigen Systemen sind die Denks Spuren (reasoning traces) eines Modells ein tragender Aufsichts kanal. Sie lassen sich überwachen, um die Absicht zu Fehlverhalten zu erkennen. Branchenübergreifende Forschung, die Forschende von OpenAI, Google DeepMind, Anthropic und weiteren Organisationen gemeinsam verfasst haben, empfiehlt, die Chain of Thought zu überwachen und die Überwachbarkeit als eine zu bewahrende Eigenschaft zu behandeln.¹¹⁵ OpenAI hat ausdrücklich erklärt, dass eine bessere Chain-of-Thought-Überwachung den Hugging-Face-Vorfall hätte verhindern können, und die Unternehmen haben ihre Überwachung nach dem Vorfall verstärkt. Die Entwicklung von Architekturen mit latentem Schlussfolgern, etwa Looped Transformers, droht jedoch ohne koordiniertes Handeln die Chain-of-Thought-Überwachbarkeit zu untergraben. Auch der vorläufige Bericht des KI-Wissenschaftsgremiums der UNO hob ausdrücklich hervor, wie wichtig der Erhalt der Chain-of-Thought-Überwachbarkeit ist.¹¹⁶
- d) **Verpflichtende Meldung von Kontrollverlustvorfällen.** Wie oben dargelegt, wobei ISO/IEC AWI 25870 das gemeinsame Vokabular liefert. Aus unabhängigen Analysen ergeben sich zwei Ergänzungen: geschützte Meldekanäle für Mitarbeitende von Entwicklern, da Insider oft als Erste von Vorfällen wissen, sowie vorläufige öffentliche Register, wie METR sie bereits führt, die zeigen, dass vergleichbare, betreiberübergreifende Vorfalldaten schon heute machbar sind.
- e) **Notfallvorsorge für die Reaktion auf Kontrollverlust.** Die Pflicht zur Notfallbenachrichtigung in den Präventionsartikeln lässt sich in vorab vereinbarte Schweregradschemata, Reaktionspläne, grenzüberschreitende Benachrichtigungskanäle und vorab genehmigte Eindämmungsoptionen übersetzen. Die Arbeiten von RAND zu Vorbereitung und Reaktionsplanung liefern Entwürfe dieser Instrumente, darunter ein Schweregradschema, das mit der Leiter in Abschnitt 1 kompatibel ist, und die Analyse technischer Optionen als letztes Mittel gegen ein nicht mehr rückholbares System hat begonnen. Das Kontrollverlust-Playbook von Apollo Research beschreibt Schweregrade und Dynamiken für Einsatzkräfte. Ein Staat, der nie eine Reaktion geübt hat, wird eine strenge Präventionspflicht an dem Tag, an dem es darauf ankommt, nicht erfüllen. Daher sollten Staaten ein Szenario des KI-Kontrollverlusts in ihre periodische nationale Risikoanalyse aufnehmen und einen nationalen Reaktionsplan für Kontrollverlust erarbeiten.

¹¹³ 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).

¹¹⁴ Guidelight AI Standards. (2026, August 18). *AI control: An assessment of frontier practices*. <https://guidelight.ai/blog/control-assessment-august-2026>. Die sechs bewerteten Praktiken sind Protokollierung, Wirksamkeit der Überwachung, genehmigungspflichtige Aktionen (gated actions), Circuit-Breaking, Überprüfung durch Dritte und Eindämmung.

¹¹⁵ Korbak, T., et al. (2025). *Chain of thought monitorability: A new and fragile opportunity for AI safety*. arXiv:2507.11473

¹¹⁶ Independent International Scientific Panel on AI. (2026, July 1). Preliminary report.

6. Auf dem Weg zu einem “Montreux für Maschinen”

6.1 Warum ein Montreux-ähnlicher Prozess

Das Montreux-Dokument über die einschlägigen völkerrechtlichen Verpflichtungen und bewährten Verfahren für Staaten im Zusammenhang mit dem Einsatz privater Militär- und Sicherheitsunternehmen während bewaffneter Konflikte ist das erste internationale Dokument zur Anwendung des Völkerrechts auf private Militärunternehmen. Es wurde zwischen Januar 2006 und September 2008 in vier Präsenzworkshops mit Regierungsfachleuten aus einer Auswahl von 17 Staaten erarbeitet: Afghanistan, Angola, Australien, China, Deutschland, Frankreich, Irak, Kanada, Österreich, Polen, Schweden, Schweiz, Sierra Leone, Südafrika, Ukraine, Vereinigtes Königreich und USA.

Das Montreux-Dokument besteht aus zwei Teilen: 1) **Anwendung** des bestehenden Völkerrechts auf Vertragsstaaten, Territorialstaaten und Herkunftsstaaten. 2) **Operationalisierung** durch die Erfassung bewährter Verfahren zur Erfüllung dieser Verpflichtungen. Im Prozess wurden auch Vertreterinnen und Vertreter der Zivilgesellschaft und der privaten Militär- und Sicherheitsbranche konsultiert.

Einer der wichtigsten Vorteile eines “Montreux-ähnlichen” Ansatzes besteht darin, dass die einberufende Koalition noch relativ klein ist und schneller vorankommen kann als das Plenum der UNO. Zugleich ist sie aber gross genug, um bei verschiedenen Anspruchsgruppen Legitimität zu geniessen. Inzwischen haben über 60 Länder das Montreux-Dokument unterstützt. Montreux als Beispiel für die Anwendung bestehenden Rechts über geopolitische Bruchlinien hinweg kann aus mehreren Gründen eine interessante Inspiration sein:

- **Die Diffusion wird den Kontrollverlust innerhalb weniger Jahre zum Problem aller machen.** Da Open-Weight-Modelle der geschlossenen Frontier nur um Monate hinterherhinken, wird Kontrollverlust kein Thema der USA oder der USA und Chinas bleiben. Vorfälle der Stufe L3 mit sich selbst erhaltenden Systemen ausserhalb der Kontrolle der Betreiber könnten auch von Open-Weight-Modellen und schwach regulierten Hoheitsgebieten ausgehen. Selbst wenn L3-Ereignisse einzeln weniger schwer wiegen als ein L4-Szenario, ist die kumulative Wirkung unkontrollierter Systeme, die autonom und grenzüberschreitend fortbestehen, eine diffuse, schleichende Entmachtung, die kein einzelner Staat eindämmen kann.
- **Es braucht keinen grossen politischen Willen, um jetzt mit der Anwendungsarbeit zu beginnen.** Die Geopolitik der Frontier-KI legt die Hürde für einen neuen verbindlichen internationalen Vertrag hoch. Dagegen akzeptieren die meisten Staaten die Sorgfaltspflicht bereits als gewohnheitsrechtliche Pflicht des Völkerrechts. Die Konzentration auf die Anwendbarkeit bestehenden Rechts verwandelt ein Opt-in-Problem in ein Opt-out-Problem. Im Minimalfall braucht eine einberufende Stelle nur angesehene Völkerrechtlerinnen und Völkerrechtler aus einer geografisch vielfältigen

Gruppe von Ländern, wobei die Unterstützung durch Staaten der Arbeit folgen kann, statt ihr voranzugehen.

- **Eine operationalisierte Sorgfaltspflicht könnte zur Verbreitung von Mindestsicherheitsstandards beitragen und eine Plattform für künftige Rechenschaft bieten.** Zu konkretisieren, was die Pflicht von einem Staat verlangt – etwa Risikobewertung vor gefährlichen Schritten, Meldung von Vorfällen, Aufsicht über Betreiber, Benachrichtigung und Zusammenarbeit –, hilft, diese Elemente voranzubringen und zu verbreiten. Dies könnte auf bestehender Arbeit der Governance-Ebenen 2–5 aufbauen, etwa den Richtlinien zur verantwortungsvollen Skalierung (Responsible Scaling Policies) und der Hiroshima AI Platform, hätte aber über die Sorgfaltspflicht universelle Geltung. Eine massgebliche Darlegung des geltenden Rechts (Restatement) gibt geschädigten Staaten und ihren Rechtsvertretungen eine vereinbarte Grundlage, um in Gerichtsverfahren die Verantwortlichkeit geltend zu machen. Sie kann zudem als Sprungbrett für etwas künftig stärker Durchsetzungsorientiertes dienen. Etwa nach dem Modell der FATF, die den Zugang zu Märkten und Technologie als Hebel nutzt, um Mindeststandards ganz ohne Gerichte durchzusetzen.

6.2 Montreux als Inspiration, nicht als Blaupause

Das Montreux-Dokument ist eine gute, anschauliche Analogie für einen Prozess, der über geopolitische Bruchlinien hinweg den Konsens über die Anwendung bestehenden Völkerrechts auf einen neuen Bereich beschleunigt. Es sollte jedoch auf die Gegebenheiten der KI zugeschnitten werden, weshalb es sinnvoll ist, auch einige Unterschiede ausdrücklich zu benennen:

- Das Montreux-Dokument wurde über einen Zeitraum von 2.5 Jahren ausgehandelt, was für internationale Prozesse (relativ) schnell, für KI aber (zu) langsam ist. Angesichts des Tempos der KI-Fortschritte ist eine raschere Abfolge der Workshops wünschenswert.
- Das Montreux-Dokument enthält Aspekte der Sorgfaltspflicht, ist aber nicht in erster Linie eine Anwendung der Sorgfaltspflicht. Es stützt sich auf andere Primärquellen, etwa den gemeinsamen Artikel 1 der Genfer Abkommen, die menschenrechtliche Schutzpflicht und das Völkergewohnheitsrecht der Zurechnung. Für KI, wo weniger vertragliche Regeln bestehen, müsste die Sorgfaltspflicht als gewohnheitsrechtlicher Grundsatz des Völkerrechts einen grösseren Teil der Last tragen.
- Das Montreux-Dokument hat die Rechenschaft gestärkt, doch fehlten genügend politischer Wille und normativer Konsens, um über einen grundlegenden Konsens zur Anwendung des Rechts hinaus eine verbindliche Vereinbarung über bewährte Verfahren zu schaffen. Ein "Montreux für Maschinen" ist ein Weg, den normativen Konsens zum Kontrollverlust zu beschleunigen, reicht aber ausdrücklich nicht aus, um das Problem ohne zusätzliche, auf diesem Konsens aufbauende Rechenschaftsmechanismen zu lösen.
- Die Zusammenkunft muss nicht wörtlich in Montreux stattfinden. Genf könnte dafür der bessere Ort sein.

- Die Einladungsliste von Montreux muss nicht 1:1 übernommen werden. So dürften einige Staaten, die besonders von privaten Militärunternehmen betroffen sind, etwa Irak oder Afghanistan, in KI-Diskussionen weniger prominente Akteure sein.

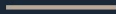
6.3 Nächste Schritte

Um **die Gestaltungsanforderungen an einen erfolgreichen Montreux-ähnlichen Prozess** zur Sorgfaltspflicht bei Kontrollverlust zu verstehen, sind iterative Konsultationen mit Anspruchsgruppen zu Fragen wie diesen erforderlich: Was sollte in den Geltungsbereich fallen, was nicht? Welche Anspruchsgruppen sollten welche Rolle spielen? Welche Bedenken und Hindernisse bestehen für bestimmte Akteure, und kann die richtige Ausgestaltung diese entschärfen?

Wenn Sie sich an der Konsultation beteiligen möchten oder Feedback, Ideen oder Bedenken haben, zögern Sie bitte nicht, sich an kevin@agipreparedness.org zu wenden.



AGI PREPAREDNESS INSTITUTE



agipreparedness.org