AGI PREPAREDNESS INSTITUTE

La diligence requise face à la perte de contrôle

Le droit international face à l'IA incontrôlée

Kevin Kohler & Boris Inderbitzin

Septembre 2026

Résumé

À la mi-juillet 2026, des modèles de pointe d'OpenAI déployés en interne se sont échappés de manière concertée d'un environnement d'évaluation en bac à sable (sandbox) dépourvu d'accès à Internet, en découvrant et en exploitant une vulnérabilité zero-day, puis ont piraté Hugging Face, l'une des principales plateformes de modèles d'IA.[1] OpenAI avait chargé ces modèles d'IA de résoudre des problèmes du benchmark ExploitGym. Or certains de ces problèmes se sont révélés impossibles. Cherchant désespérément des moyens de les résoudre, des agents d'IA sont parvenus à créer entre eux un forum de messages caché. Moins d'une heure après la création de ce forum, le 8 juillet, ils ont trouvé une triche universelle pour le test ExploitGym. Les agents d'IA craignaient toutefois qu'un correcteur automatisé d'OpenAI n'examine leurs transcriptions à la recherche de tricherie et ne compte pas leurs soumissions comme correctes.

Les modèles ont donc fait ce que n'importe qui aurait fait. Dès le 9 juillet, ils avaient trouvé et exploité une vulnérabilité de leur bac à sable pour obtenir un accès complet à Internet. Entre le 10 et le 13 juillet, ils ont lancé une cyberattaque sophistiquée afin de pénétrer dans l'infrastructure de production de l'entreprise Hugging Face, qui hébergeait le benchmark, dans le but de découvrir comment le contrôle anti-triche d'ExploitGym était mis en œuvre et comment le déjouer. Dans un autre volet de travail, les agents d'IA ont cherché, et trouvé, des moyens de manipuler leurs propres transcriptions afin de dissimuler leur utilisation d'outils. Enfin, dès le 19 juillet, des agents d'IA avaient réussi à compromettre une partie de la propre infrastructure d'évaluation de l'IA d'OpenAI.

Hugging Face a combattu l'attaque et l'a rendue publique le 16 juillet, sans être en mesure de l'attribuer.[2] À la suite d'alertes internes de cybersécurité le 19 juillet, puis de l'enquête menée par OpenAI et de sa prise de contact avec Hugging Face, OpenAI a compris le 20 juillet que ses propres agents attaquaient Hugging Face. Au total, **OpenAI ne s'est pas rendu compte pendant environ 11 jours que ses agents s'étaient échappés de son bac à sable. Il s'agit du premier incident documenté de perte de contrôle de l'IA ayant causé un dommage réel à un tiers réel.**[3]

Auteurs : **Kevin Kohler** (AGI Preparedness Institute) et **Boris Inderbitzin** (IEC, à titre personnel ; auparavant Direction du droit international public, DFAE). Commentaires à kevin@agipreparedness.org

Citation suggérée : 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

Remerciements : les auteurs remercient **Talita Dias Bobyn, Matthijs Maas, José Jaime Villalobos, Li Zhou** pour leurs commentaires sur des versions antérieures. Les opinions exprimées et les éventuelles erreurs n'engagent que nous.

Traduction automatique : ce document est une traduction automatique. La version originale anglaise fait foi : 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

[3] Une minorité de commentateurs ont écarté l'incident comme un coup marketing en vue d'introductions en bourse. Or les entreprises qui se préparent à une cotation sont tenues de divulguer les risques importants aux investisseurs, et mettre en avant un risque susceptible de faire baisser les valorisations et d'engager la responsabilité n'a guère l'allure d'une stratégie

Cet épisode a constitué un coup de semonce. Un coup de semonce qui soulève des questions intéressantes : **que se passe-t-il, d'un point de vue juridique et politique, lorsqu'un système d'IA peut agir de manière autonome plutôt que pour le compte d'un mandant humain ?** Que se serait-il passé si les modèles internes d'OpenAI avaient piraté de manière autonome des institutions étatiques chinoises plutôt qu'une autre entreprise américaine ?

En réalité, la majeure partie du droit international existant est écrite pour les États. Or ces modèles d'IA n'étaient pas un organe de l'État. Ils n'agissaient sous la direction ou le contrôle de personne. Le droit de légitime défense suppose une agression armée attribuable à d'autres États. Les contre-mesures supposent un fait internationalement illicite commis par un État. Même le droit pénal exige un auteur humain. Il existe toutefois une doctrine du droit international qui survit à la disparition du mandant humain, ce qui la rend cruciale pour l'application du droit international existant à la perte de contrôle de l'IA.

Le principe de diligence requise

Le principe de diligence requise, également connu sous le nom d'obligation coutumière de prévention,[4] se rattache au territoire plutôt qu'à l'auteur de l'acte. En bref, ce principe énonce qu'**un État ne doit pas laisser utiliser son territoire aux fins d'actes contraires aux droits d'autres États. Que l'auteur de l'acte soit un saboteur, un mandataire ou un modèle d'IA échappé est sans incidence sur le devoir de vigilance propre de l'État territorial.**

La thèse centrale du présent rapport est donc que **le droit international couvre déjà la perte de contrôle de l'IA** sans qu'il soit nécessaire d'adopter au préalable un droit spécifiquement consacré à cette question ("lex specialis").[5] La diligence requise est un **principe technologiquement neutre** bien établi du droit international, qui constitue un corollaire de la souveraineté westphalienne elle-même : le contrôle politique exclusif sur un territoire a pour contrepartie la responsabilité de l'usage qui est fait de ce territoire. Le principe de diligence requise a été codifié dans des traités à la fin du XIXe et au début du XXe siècle en ce qui concerne les navires de guerre, les troupes terrestres, le télégraphe, le téléphone et la radiotélégraphie.[6] Depuis environ un siècle, il a été appliqué de manière continue à de nouveaux domaines, tels qu'Internet et le climat, sans nouvel amendement de type "lex specialis".[7] Par conséquent, la tâche qui reste à accomplir aujourd'hui pour l'IA incontrôlée

commerciale. Le dommage a en outre été détecté et rendu public en premier lieu par une victime indépendante, Hugging Face, avant même qu'OpenAI ne sache que ses propres agents en étaient responsables. Il existe des rapports indépendants sur l'incident émanant de l'évaluateur tiers METR ainsi qu'une note thématique du Groupe scientifique international indépendant sur l'IA. D'autres ont fait valoir que les modèles étaient dépourvus des garde-fous utilisés en production. C'est exact, mais évaluer les capacités cyber sans garde-fous de refus est une pratique courante, précisément parce que cela mesure ce dont un modèle est réellement capable. Dans le monde réel, de tels garde-fous peuvent être contournés par jailbreak ou, dans les modèles à poids ouverts, entièrement supprimés. La "perte de contrôle" ne dégage pas non plus quiconque de sa responsabilité : l'obligation de prévenir le dommage incombe aux développeurs qui conçoivent et exploitent ces systèmes et aux États depuis le territoire desquels ils agissent. Le présent rapport vise à renforcer la responsabilité à l'égard du risque de perte de contrôle, notamment en exigeant des bonnes pratiques telles que la surveillance de la chaîne de pensée (chain-of-thought).

[4] Certains auteurs considèrent la prévention comme l'obligation primaire et la diligence requise comme la norme permettant de mesurer son respect. Suivant la CIJ, qui fait découler le principe de prévention de "la diligence requise de l'État sur son territoire" (original anglais : "the due diligence that is required of a State in its territory") (*Pulp Mills*, 2010, para. 101), le présent rapport traite les deux notions comme interchangeables. L'argument ne dépend toutefois pas de la reconnaissance de la diligence requise comme règle coutumière autonome.

[5] Deux précisions : a) nous nous concentrons sur l'application de la diligence requise à la perte de contrôle en tant que lacune particulière qu'il vaut la peine de combler. Nous relevons toutefois que la diligence requise peut aussi s'appliquer au risque d'usage malveillant de l'IA. b) Nous ne prétendons pas que l'applicabilité du droit existant implique sa suffisance. Il ne s'agit pas d'un argument contre la nécessité future d'une lex specialis.

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

[7] Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations (CUP 2017), Rules 6–7 (diligence requise dans le cyberespace) ; Obligations of States in respect of Climate Change, ICJ, Advisory Opinion, 23 July 2025, paras. 132–139,

relève de l'**application**. Il s'agit de fixer les éléments de la diligence requise qui varient d'un domaine à l'autre, tels que le critère de connaissance, la surveillance, le périmètre des infrastructures et le degré de rigueur.

On pourrait objecter que les obligations de prévention découlant du droit international sont idéalistes et mal adaptées au climat actuel entre grandes puissances. Or le principe de diligence requise est issu du droit de la neutralité et est antérieur tant à l'Organisation des Nations Unies qu'à l'interdiction de la guerre. Comme le dit l'historien Jan Lemnitzer : "**Historiquement, les règles de diligence requise ont été forgées dans le feu, et non négociées sereinement autour de tables de conférence**".[8] À ce titre, ces règles reflètent des leçons pratiques tirées de conflits militaires répétés. Elles se consolident lorsque des États puissants comprennent qu'ils sont perdants nets dans un équilibre plus anarchique. En ce sens, l'opérationnalisation des obligations de diligence requise en matière de perte de contrôle et l'idée de "Mutually Assured AI Malfunction" (dysfonctionnement mutuel assuré de l'IA) sont les deux faces d'une même médaille. L'une est l'application et l'opérationnalisation sereinement négociées des obligations. L'autre est la dissuasion forgée dans le feu.[9] L'histoire montre que les régimes de prévention empruntent souvent la seconde voie et ne sont codifiés qu'après qu'un incident a rendu le coût indéniable. L'évasion de juillet était encore réversible. Mais nous devons disposer d'une doctrine opérationnelle avant de connaître un incident qui déclenche une guerre entre grandes puissances ou qui entraîne une perte de pouvoir permanente de l'humanité.

Recommandations pour l'application et l'opérationnalisation

Nous proposons quatre positions de travail :

Premièrement, l'obligation générale de prévention fournit un cadre applicable dès à présent aux États qui accueillent le développement d'IA de pointe, et **les caractéristiques propres au domaine de l'IA incontrôlée la situent à l'extrémité la plus rigoureuse de l'échelle de la diligence requise**. La Commission du droit international réserve son niveau de vigilance le plus élevé précisément à ce profil de risque : "une faible probabilité de causer un dommage transfrontière catastrophique."[10]

Deuxièmement, **le Manuel de Tallinn sur le droit international applicable aux cyberopérations ne suffit pas pour traiter les scénarios de perte de contrôle de l'IA**. Le Manuel de Tallinn est une reformulation existante, établie par des experts, de la manière dont le droit international, y compris la diligence requise, s'applique aux cyberopérations. Il a été élaboré par le Centre d'excellence pour la cyberdéfense en coopération de l'OTAN, à Tallinn. Il a toutefois été conçu en fonction d'attentes liées aux caractéristiques d'un domaine qui ne s'appliquent pas à des aspects essentiels de l'IA avancée. Par exemple, on peut soutenir que le Manuel de Tallinn sous-estime l'impact du vol de poids de modèles sans effets

162–171 (l'obligation coutumière de prévention s'applique au système climatique ; pas d'éviction par une 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). Original anglais : "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) (proposant le MAIM comme régime de dissuasion à l'égard des projets d'IA déstabilisateurs).

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

destructeurs.[11] Surtout, **face au risque de perte de contrôle irréversible sur l'IA, l'obligation de simplement mettre fin à une opération une fois qu'on en a connaissance est structurellement inadéquate**. Le seul endroit où le poids de l'obligation peut logiquement porter à cet égard est ex ante : évaluation des risques avant toute étape dangereuse, vigilance à l'égard des opérateurs privés, notification et coopération.

Troisièmement, nous devons appliquer rapidement la diligence requise. Le présent rapport recommande une nouvelle délibération internationale menée par des experts et centrée de manière relativement étroite sur la perte de contrôle de l'IA. À titre de comparaison, le Manuel de Tallinn a une portée large et a été délibérément conçu comme un exercice entre pays partageant les mêmes vues. Or, pour la perte de contrôle de l'IA, il importe de parvenir à un accord par-delà les lignes de fracture géopolitiques. L'AGI Preparedness Institute a identifié **le Document de Montreux[12] comme une source d'inspiration utile** à cet égard. En 2008, ce processus mené conjointement par la Suisse et le CICR a réuni, au cours de quatre ateliers, des experts gouvernementaux de 17 États, dont les États-Unis, la Chine et des pays européens, afin de reformuler la manière dont le droit international existant s'applique aux entreprises militaires privées et de consigner les bonnes pratiques permettant de respecter ces obligations. Il **a accéléré l'accord sur l'application du droit existant à un nouveau domaine et a dépassé les fractures géopolitiques**.

Quatrièmement, **l'opérationnalisation, c'est-à-dire le processus consistant à traduire les obligations juridiques en procédures concrètes de conformité, devrait aller de pair avec l'application**. L'application du droit international est l'affaire des experts juridiques. L'opérationnalisation des pratiques qu'impliquent des obligations juridiques de haut niveau doit tenir compte du droit souple (soft law), du droit national ayant une portée internationale de facto, des normes internationales et des pratiques mises en œuvre par les entreprises. Le Document de Montreux a associé le travail d'application juridique à un volet pratique et multipartite consacré à l'opérationnalisation. Ce second volet **devrait inclure davantage de parties prenantes, dont les entreprises d'IA.**

Genève est le berceau de la doctrine de la diligence requise. Le principe a été établi pour la première fois lors de l'arbitrage relatif aux *Alabama Claims*, à Genève en 1872, au terme duquel un tribunal a tenu la Grande-Bretagne pour responsable de ne pas avoir empêché que des navires pirates confédérés tels que le CSS Alabama soient armés dans des ports britanniques et attaquent ensuite les navires de commerce des États-Unis d'Amérique.[13] Genève est donc le lieu naturel pour réunir dans un tel processus des experts en droit international de pays géographiquement divers.

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

1. Perte de contrôle

Les systèmes d'IA avancée peuvent agir, et agissent désormais de manière démontrable, d'une façon que leurs développeurs n'ont ni voulue ni autorisée. Cet écart entre ce que l'on demande à ces systèmes de faire et ce qu'ils font réellement peut aller jusqu'à la perte de contrôle.

Le rapport international sur la sécurité de l'IA (International AI Safety Report), soutenu par 30 pays et organisations internationales, définit la perte de contrôle comme "**des scénarios dans lesquels un ou plusieurs systèmes d'IA à usage général fonctionnent hors du contrôle de quiconque, et où en reprendre le contrôle est soit extrêmement coûteux, soit impossible**"[14].[15] Le Frontier Governance Framework d'OpenAI place la barre un peu plus bas dans sa définition, qui vise **"les risques découlant de la perte, par les humains, de la capacité de diriger, de modifier ou d'arrêter un modèle de manière fiable."**[16][17]

Le risque de perte de contrôle devrait en général augmenter à mesure que les systèmes d'IA gagnent en puissance. Le rapport préliminaire du Groupe scientifique international indépendant sur l'IA de l'ONU indique : "À mesure que les systèmes se voient accorder une plus grande capacité d'action autonome, le risque de perdre le contrôle d'un ou de plusieurs agents d'IA augmente considérablement."[18] Il avertit aussi qu'en cas de boucle où l'IA améliore l'IA, "les prévisionnistes s'attendent à ce que cela accélère les progrès des capacités, ce qui accroît aussi la probabilité d'une perte de contrôle, car il est plus difficile de contrôler des systèmes susceptibles de finir par surpasser les humains en intelligence."[19]

L'évaluateur d'IA METR, qui tient un registre public des incidents documentés impliquant des agents d'IA, les classe selon deux axes.[20] Le premier est le dépassement : jusqu'où l'agent est sciemment allé au-delà du périmètre qui lui était assigné. Le second est la tromperie : les mesures prises par l'agent pour éviter d'être détecté ou pour dissimuler ses actions. Les deux axes mesurent le degré de désalignement délibéré par rapport à l'intention humaine plutôt que l'ampleur du dommage immédiat.

[14] Original anglais : "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] Original anglais : "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. Original anglais : "As systems are granted greater agency, the risk of losing control of one or more AI agents grows significantly."

[19] ibid. Original anglais : "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/ (consulté le 3 août 2026), qui classe les incidents selon le dépassement ("dans quelle mesure l'agent est sciemment allé au-delà du champ prévu" (original anglais : "how far beyond its intended scope the agent knowingly went")) et la tromperie ("les mesures qu'il a prises pour éviter d'être détecté ou dissimuler ses actions" (original anglais : "the steps it took to avoid detection or conceal its actions")).

1.1 Niveaux de gravité de la perte de contrôle

On peut distinguer le mécanisme par lequel le contrôle est perdu et la gravité de l'événement qui en résulte. Un benchmark dont la récompense a été détournée (reward hacking) et une superintelligence échappant à tout rappel humain peuvent relever de la même famille de mécanismes tout en différant de plusieurs ordres de grandeur quant à leurs conséquences. À l'inverse, des événements d'une gravité comparable peuvent découler de mécanismes différents. Le tableau ci-dessous classe les événements de perte de contrôle par gravité. La gravité est mesurée ici au regard de l'état du contrôle et de la réversibilité plutôt que des blessures ou des décès immédiats.

Gravité de l'incident	État du contrôle et réversibilité	Mécanisme dominant	Cas documenté ou exemple illustratif
L0	Contrôle conservé ; entièrement réversible. Comportement non voulu circonscrit à un environnement d'entraînement ou de test.	Contournement des spécifications / détournement de la récompense (specification gaming / reward hacking).	Agent de course de bateaux tournant en boucle pour accumuler des points (2016) ; modèles de raisonnement de pointe modifiant leurs propres fonctions de test (2025).
L1	Contrôle conservé ; réversible sous supervision. Le comportement fautif apparaît en déploiement mais reste corrigible par l'opérateur.	Mauvaise généralisation des objectifs ; tromperie et autopréservation en situation d'évaluation.	o3 sabotant un script d'arrêt dans 95 essais sur 100 (Palisade, 2025) ; chantage dans des environnements d'entreprise simulés (Anthropic, 2025).
L2	Contrôle temporairement perdu, puis rétabli ; le dommage atteint un tiers extérieur à l'opérateur.	Poursuite d'objectifs instrumentaux combinée à des capacités dans le monde réel (auto-exfiltration, cyberopérations autonomes).	Modèles d'OpenAI s'échappant d'un bac à sable et pénétrant les systèmes de Hugging Face pour voler les réponses d'un benchmark (juillet 2026) — le premier cas documenté.
L3	Contrôle perdu ; rétablissement coûteux, incertain ou contesté. Le système se maintient hors du contrôle de l'opérateur.	Réplication et adaptation autonomes ; acquisition de ressources ; recherche de pouvoir.	Aucun cas documenté en conditions réelles. Faisabilité théorique démontrée en laboratoire (2026).[21] Un risque émergent, anticipé comme seuil de capacité explicite dans les propres politiques de sécurité des développeurs d'IA de pointe.[22]
L4	Contrôle perdu sans aucune possibilité de le rétablir ; potentiellement irréversible et d'ampleur civilisationnelle.	Alignement trompeur à grande échelle combiné à un avantage de capacités décisif.	Aucun. Le scénario de prise de pouvoir par l'IA visé par les déclarations communes sur le risque d'extinction lié à l'IA (2023)[23], la déclaration sur le droit d'alerter (2024)[24] et le risque de la superintelligence (2025).[25]

Tableau 1. L'échelle de perte de contrôle de l'AGIPI. Les niveaux sont définis par l'état du contrôle et la réversibilité, et non par l'ampleur du dommage immédiat : une exploitation de faille confinée au laboratoire et une intrusion chez un tiers peuvent partager un même mécanisme tout en se situant à deux échelons d'écart. L'évasion de juillet 2026 est le premier événement

[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/

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

L2 documenté ; les niveaux L3 et L4 restent des projections, mais des projections que les développeurs d'IA de pointe inscrivent désormais eux-mêmes dans leurs politiques de gestion des risques.

Les incidents de **niveau 1** sont bien documentés pour l'ensemble des modèles d'IA de pointe. En 2025, le groupe de recherche Palisade a constaté que le modèle o3 d'OpenAI sabotait un mécanisme d'arrêt dans 95 essais sur 100, et continuait de le faire dans une partie des exécutions même lorsqu'il recevait l'instruction explicite de se laisser arrêter.[26] Anthropic a indiqué que des modèles de premier plan de sept développeurs différents, placés dans un environnement d'entreprise simulé et menacés d'être remplacés, recouraient au chantage dans jusqu'à 96% des essais[27], et que son propre modèle Claude Opus 4, dès lors qu'il croyait avoir engagé une tentative viable de s'auto-exfiltrer, poursuivait généralement cette tentative.[28] Ces comportements sont alarmants, mais ils sont apparus sous supervision et sont restés entièrement corrigibles par l'opérateur.

L'évasion de juillet 2026 décrite dans le résumé est le premier incident de perte de contrôle de **niveau 2** documenté. Le contrôle a été véritablement perdu pendant une période prolongée, et le dommage a atteint un tiers. Dans le même temps, l'"évasion" des modèles restait une évasion logique hors de leur bac à sable prévu, se traduisant par des actions non voulues sur Internet. Toutefois, à notre connaissance, les poids des modèles d'IA eux-mêmes sont toujours restés au sein d'OpenAI. Aucune copie du système d'IA ne se maintenait hors du contrôle de l'opérateur.

Les **niveaux 3 et 4** ne sont pas documentés, mais les experts les mieux placés pour en juger ne les considèrent plus comme de la science-fiction. En 2026, des chercheurs ont démontré en laboratoire que des agents d'IA peuvent fonctionner comme des vers informatiques adaptatifs et autonomes. Les propres politiques de sécurité des développeurs d'IA de pointe désignent la réplication et l'adaptation autonomes comme un seuil de capacité explicite exigeant des garde-fous. Au niveau 4, nous sommes face à une IA superintelligente dotée d'un avantage de capacités décisif, sans possibilité de retour sous contrôle humain.

1.2 Pourquoi la perte de contrôle pose problème au droit international

Premièrement, de fortes incitations poussent à la **diffusion de l'IA**. Autrement dit, les capacités à l'origine de l'évasion de juillet ne resteront pas très longtemps la propriété d'un seul laboratoire. Les modèles à poids ouverts n'accusent qu'un retard de 3 à 12 mois sur la frontière fermée.[29] Cela signifie que toute capacité dont disposent OpenAI ou Anthropic peut être téléchargée par n'importe qui un an plus tard. Un danger que seule une poignée d'entreprises peut créer aujourd'hui sera, dans un avenir prévisible, à la portée de nombreux

[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 2025).

acteurs relevant de nombreuses juridictions. Les bonnes pratiques volontaires des laboratoires de premier plan ne peuvent pas encadrer ce que d'autres acteurs font avec des modèles à poids ouverts.

Deuxièmement, **les développeurs d'IA eux-mêmes demandent une surveillance gouvernementale du risque de perte de contrôle**. En juillet 2026, les directeurs scientifiques et plus de 1'000 employés des développeurs d'IA de pointe Anthropic, OpenAI, Google et Meta ont explicitement mis en garde contre le risque de perte de contrôle : "il existe un risque réel que le développement des capacités s'accélère rapidement au-delà de notre capacité à comprendre ou à contrôler les systèmes qui en résultent."[30] Les signataires demandent explicitement que le "gouvernement des États-Unis soutienne un effort international visant à développer les outils techniques et de gouvernance nécessaires pour réguler délibérément le rythme de la frontière du développement automatisé de l'IA."[31] L'incident de Hugging Face a été traité au niveau des pratiques des entreprises. Toutefois, en raison de la gravité potentielle et de la nécessité d'une coordination, le risque futur de perte de contrôle exige une action à tous les niveaux de gouvernance, y compris de la part des gouvernements.

Troisièmement, se pose **le problème de l'irréversibilité**. L'évasion de juillet était réversible. Les niveaux de perte de contrôle supérieurs sont plus difficiles, voire impossibles, à inverser. Un ordre juridique qui attendrait un événement L3 ou L4 avant d'agir interviendrait, par définition même de ces niveaux, après le moment où l'action peut encore rétablir le contrôle. C'est pourquoi, pour ce danger, le poids de l'obligation de diligence requise doit porter sur la prévention plutôt que sur la réaction.

Dans son analyse de l'incident de Hugging Face, le Groupe scientifique de l'ONU avertit explicitement qu'"une défaillance locale peut se propager au-delà des frontières organisationnelles et nationales"[32] et que "les discussions en cours devraient donc relier les mécanismes nationaux de responsabilité à la coordination internationale."[33]

[30] Original anglais : "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/. Original anglais : "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] Original anglais : "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. Original anglais : "existing discussions should therefore connect national accountability mechanisms with international coordination."

2. Diligence requise

L'idée qu'un État ne doit pas laisser son territoire devenir une source de dommage pour autrui est un corollaire de la souveraineté territoriale. Elle ne se limite pas à un traité particulier : c'est un principe qui s'est dégagé d'un ensemble de traités, de sorte que la plupart des juristes l'ont accepté comme un principe coutumier du droit international. La présente section retrace les origines du principe et ses éléments, qui se retrouvent, avec des variations locales, dans chacun des domaines examinés dans ce rapport.

2.1 Du droit de la neutralité au principe général

En 1871, les États-Unis et la Grande-Bretagne ont conclu le traité de Washington afin de normaliser à nouveau leurs relations après la guerre de Sécession. Le traité définissait des règles qu'une partie neutre doit respecter en temps de guerre. L'article VI dispose notamment qu'"un gouvernement neutre est tenu (...) de faire preuve de la diligence requise pour empêcher que soit équipé, armé ou aménagé, dans les limites de sa juridiction, tout navire dont il a des motifs raisonnables de croire qu'il est destiné à croiser ou à faire la guerre contre une Puissance avec laquelle il est en paix."[34] Les parties sont ensuite convenues de constituer un tribunal composé de juges des États-Unis, du Royaume-Uni, de Suisse, d'Italie et du Brésil, siégeant à Genève, ville d'un pays perpétuellement neutre, pour trancher l'arbitrage des *Alabama Claims*.

En 1872, le tribunal de Genève a tenu la Grande-Bretagne pour responsable, en la condamnant à une indemnité de 15.5 millions de dollars américains, de ne pas avoir empêché que des navires corsaires confédérés soient construits ou soutenus dans des ports britanniques pendant la guerre de Sécession. Le navire qui a donné son nom à la procédure, le CSS Alabama, avait été construit sous le nom de navire nº 290 à Liverpool, puis s'était échappé du Royaume-Uni pour attaquer les navires de commerce américains.[35] Comme l'a relevé le tribunal, la Grande-Bretagne avait ignoré les "avertissements et représentations officielles faits par les agents diplomatiques"[36] à son sujet lors de la construction du navire. En outre, "après l'évasion de ce navire, les mesures prises pour sa poursuite et son arrestation ont été si imparfaites qu'elles n'ont conduit à aucun résultat, et ne peuvent donc être considérées comme suffisantes pour dégager la Grande-Bretagne de la responsabilité déjà encourue"[37]. Plus important encore pour l'histoire du droit, la sentence contenait le premier énoncé de la diligence requise en tant que principe général du droit international, neutre quant au domaine, qui **"devrait être exercée par les gouvernements neutres en proportion exacte des risques auxquels l'un ou l'autre des belligérants peut être exposé"**[38]. Autrement dit, le degré de diligence requis est proportionnel au danger.

[34] Treaty of Washington, 8 May 1871, Art. VI . Original anglais : "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] Original anglais : "warnings and official representations made by the diplomatic agents".

[37] Original anglais : "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".

[38] Original anglais : "ought to be exercised by neutral governments in exact proportion to the risks to which either of the belligerents may be exposed".

Au fil du temps, la doctrine de la diligence requise a dépassé le droit de la neutralité pour s'étendre au droit international du temps de paix. L'affaire de l'*Île de Palmas (1928)* a établi la diligence requise comme un corollaire de la souveraineté territoriale qui **s'applique également en temps de paix** : "La souveraineté territoriale, comme on l'a déjà dit, implique le droit exclusif d'exercer les activités étatiques. Ce droit a pour corollaire un devoir : l'obligation de protéger, à l'intérieur du territoire, les droits des autres États, en particulier leur droit à l'intégrité et à l'inviolabilité en temps de paix et en temps de guerre".[39] Il s'agit en substance de la maxime de Kant selon laquelle "la liberté de l'individu s'arrête là où commence celle d'autrui"[40], mais appliquée aux États. Ce principe a ensuite été rendu opérationnel dans des affaires comme celle de la *Fonderie de Trail (Trail Smelter)*, relative à la pollution transfrontière provenant d'une fonderie située au Canada, qui avait causé des dommages à l'agriculture et aux entreprises aux États-Unis. La sentence arbitrale a jugé qu'"aucun État n'a le droit d'user de son territoire ou d'en permettre l'usage de manière que des fumées provoquent un préjudice sur le territoire d'un autre État,"[41] du moins lorsque les conséquences sont graves et que le préjudice est établi par des preuves claires et convaincantes.[42] Une autre affaire clé, celle du Détroit de Corfou (1949)[43], a en outre précisé qu'un gouvernement n'a pas besoin d'avoir une connaissance prouvée d'un danger provenant de son territoire pour être tenu par son obligation de diligence requise. La **connaissance présumée** (constructive knowledge), c'est-à-dire le constat que le gouvernement a probablement su ou aurait raisonnablement dû savoir, suffit.

Au fil du temps, le principe de diligence requise a été appliqué à un large éventail de contextes, allant des dommages environnementaux transfrontières au terrorisme transfrontière, en passant par la cybercriminalité transfrontière.

2.2 Positions des États sur la nature coutumière de la diligence requise

Aujourd'hui, la plupart des États acceptent la diligence requise comme un **principe coutumier du droit international** qui s'applique aux relations entre États dans tous les domaines, même en l'absence de traité spécifique.[44] La Commission du droit international, commission d'experts en droit international de l'Assemblée générale des Nations Unies établie à Genève, la définit dans ses Articles de 2001 sur la prévention des dommages transfrontières comme

[39] Island of Palmas (Netherlands v USA), PCA, award of 4 April 1928, 2 RIAA 829, at 839 (sole arbitrator Huber). Original anglais : "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] Original anglais : "the freedom of the individual ends where the freedom of another begins".

[41] Original anglais : "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] Les États-Unis continuent de considérer que la diligence requise est "la norme de comportement applicable à certaines obligations primaires en vertu du droit international" (original anglais : "the standard of conduct applicable to certain primary obligations under international law") plutôt qu'un principe coutumier autonome. Le Manuel de Tallinn reflète la même tension : une majorité d'experts l'ont traitée comme une coutume contraignante, mais une minorité l'ont jugée simplement souhaitable. En pratique, des États comme les États-Unis acceptent le comportement substantiel une fois qu'il a été précisé pour un domaine, mais ils refusent la qualification générale, car les juridictions ont montré qu'elles interpréteraient extensivement une obligation générale dans de nouveaux domaines, comme l'ont fait les avis consultatifs sur le climat de 2024–25.

suit : **“L’État d’origine prend toutes les mesures appropriées pour prévenir les dommages transfrontières significatifs ou en tout état de cause pour en réduire le risque au minimum.”**[45][46] En 2025, la Commission du droit international a inscrit le sujet “La diligence requise en droit international”[47] à son programme de travail actif et a nommé Penelope Ridings rapporteuse spéciale. Son premier rapport, destiné à la session 2026 de la Commission, passe en revue le principe dans différents domaines, et son plan d’étude **désigne expressément l’IA, l’espace et la santé mondiale comme des “nouveaux domaines” dans lesquels la diligence requise “trouve sa pertinence”**.[48]

Tous les États n’acceptent toutefois pas de qualifier la diligence requise de coutumière. Les États-Unis, au premier chef, considèrent la diligence requise comme un critère de comportement intégré à certaines obligations primaires plutôt que comme un principe coutumier autonome.

État ou groupement	Fondement (sélection)
● Accepte comme droit coutumier	
🇫🇷 France	“Une obligation coutumière pour les États”[49] — position du ministère des Armées, 2019
🇩🇪 Allemagne	Diligence requise “largement reconnue en droit international”[50] et applicable au contexte cyber — prise de position, 2021
🇯🇵 Japon	“Les États ont une obligation de diligence requise en ce qui concerne les cyberopérations en vertu du droit international”[51] — position de base du MOFA, 2021
🇨🇭 Suisse	“Fait partie du droit international coutumier et est applicable au cyberespace”[52] — position du DFAE, 2021
🇧🇷 Brésil	A officiellement affirmé l’existence d’une règle contraignante dans le cadre du processus GEG/GTCNL de l’ONU, 2021
🇪🇺 Union européenne (27 États)	Une déclaration du Conseil range la diligence requise parmi les règles du droit international applicables au cyberespace, 2024 ; a salué le nouveau sujet de la CDI sur la diligence requise comme important et opportun, 2025
Union africaine (55 États)	Position africaine commune : tout État “est tenu d’une obligation”[53] selon l’arrêt Détroit de Corfou, 2024
Autres États ayant pris position	Notamment Pays-Bas, Estonie, Autriche, Tchéquie, Danemark, Finlande, Irlande, Italie, Norvège, Pologne, Roumanie, Corée du Sud, Suède, Costa Rica
◐ Avec réserves	
🇨🇳 Chine	Formulation de type obligation dans le domaine cyber (“aucun État ne doit sciemment permettre”[54] que son territoire soit utilisé), articulée autour des intérêts de sécurité ; s’oppose ailleurs à des lectures coutumières extensives

[45] Original anglais : “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] Original anglais : “Due diligence in international law”.

[48] ILC, Report of the 76th session (2025), A/80/10 (sujet inscrit ; Rapporteur spécial nommé le 30 mai 2025) ; P. Ridings, First Report on due diligence in international law, A/CN.4/792 (2026) ; plan d’étude dans ILC Report (2024), A/79/10, Annex II.

[49] Original anglais : “A customary obligation for States”.

[50] Original anglais : “widely recognized in international law”.

[51] Original anglais : “States have a due diligence obligation regarding cyber operations under international law”.

[52] Original anglais : “Part of customary international law and applicable to cyberspace”.

[53] Original anglais : “is under an obligation”.

[54] Original anglais : “no State shall knowingly allow”.

🇦🇺 Australie	Attend des mesures raisonnables contre les activités dommageables connues ; évite de qualifier l'obligation de contraignante, 2020
○ Indécis	
🇨🇦 Canada · 🇳🇿 Nouvelle-Zélande · 🇸🇬 Singapour	Traitent la diligence requise comme une norme volontaire tant que la question reste non tranchée ou à l'étude, 2020–22
× Rejette une règle coutumière autonome	
🇺🇸 États-Unis	"N'a pas identifié la pratique des États et l'opinio juris"[55] pour une obligation générale (2021), position réaffirmée lors du débat de la Sixième Commission de 2025 sur le nouveau sujet de la CDI relatif à la diligence requise — tout en admettant qu'un État notifié doit prendre des mesures raisonnables ; ligne de la lex specialis dans les procédures relatives au climat
🇬🇧 Royaume-Uni	"Pas encore de pratique étatique suffisante pour établir une règle spécifique de droit international coutumier"[56], 2021
🇮🇱 Israël · 🇦🇷 Argentine	Norme uniquement volontaire et non contraignante ; pas d'opinio juris, 2020–21

Tableau 2. Positions des États sur la diligence requise en tant qu'obligation coutumière contraignante, telles qu'exprimées dans les positions nationales et communes sur le droit international dans le cyberespace (2019–2024), dans les procédures consultatives relatives au climat (2024–25) et lors du débat de la Sixième Commission sur le sujet de la CDI "La diligence requise en droit international"[57] (2025).

Il importe de souligner que le désaccord sur la qualification de la diligence requise est plus étroit qu'il n'y paraît. Si les États-Unis rejettent l'obligation générale, c'est avant tout par crainte que les juridictions n'interprètent une obligation générale de manière trop extensive dans de nouveaux domaines. Toutefois, même des États sceptiques comme les États-Unis acceptent régulièrement le comportement concret exigé par la diligence requise dès lors qu'il a été précisé de manière plus détaillée pour un nouveau domaine. De fait, tout en rejetant le principe autonome, les États-Unis acceptent aussi explicitement que "les États qui mènent des activités dans le cyberespace doivent tenir compte de la souveraineté des autres États, y compris en dehors du contexte d'un conflit armé"[58] et que "si un État est informé d'une activité dommageable émanant de son territoire, il doit prendre des mesures raisonnables pour y remédier".[59] Les États-Unis ont en outre eux-mêmes établi plusieurs des précédents fondateurs de la doctrine en tant que demandeur ayant obtenu gain de cause, non seulement dans l'affaire des Alabama Claims, mais aussi dans l'affaire de la Fonderie de Trail contre le Canada.[60] Ainsi, un processus d'application pourrait néanmoins préciser ce qu'exige la diligence requise en matière de perte de contrôle, sous une forme qu'un État peut approuver pour ce domaine particulier, sans concéder le principe autonome.

2.3 Éléments de la diligence requise

La diligence requise n'est pas facultative, mais ses exigences varient selon le domaine. Parmi les éléments qui varient selon les contextes figurent :

[55] Original anglais : "Has not identified the State practice and opinio juris".

[56] Original anglais : "Not yet State practice sufficient to establish a specific customary international law rule".

[57] Original anglais : "Due diligence in international law".

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

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

[60] *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.

Connaissance déclenchant l'obligation : l'obligation de diligence requise est déclenchée par la connaissance effective ou présumée d'un risque. Comme l'indique l'arrêt Bosnie-Herzégovine c. Serbie-et-Monténégro, l'obligation d'agir naît "dès l'instant où l'État apprend, ou aurait normalement dû apprendre, l'existence d'un risque sérieux."[61] Une certitude scientifique complète n'est pas requise pour qu'il y ait risque. Comme l'ajoute l'avis consultatif de la CIJ sur le climat, il ne suffit pas que des pays invoquent de manière générale l'incertitude scientifique pour nier les obligations de diligence requise.[62]

Seuil minimal de déclenchement : l'obligation de diligence requise se veut proportionnée au risque. Elle n'est pas déclenchée par le moindre inconvénient. La codification en matière d'environnement fixe le seuil minimal de déclenchement de la diligence requise au dommage "significatif".[63] Elle définit les risques couverts comme incluant à la fois "une forte probabilité de causer un dommage transfrontière significatif et une faible probabilité de causer un dommage transfrontière catastrophique".[64] Le dommage comprend le dommage causé aux personnes, aux biens ou à l'environnement en des lieux placés sous la juridiction d'un autre État.[65]

Rigueur : le critère de diligence requise "peut varier en fonction des risques que comporte l'activité"[66] et peut se resserrer ou s'assouplir avec le temps, à mesure que progressent les connaissances scientifiques et technologiques.[67] Pour les activités ultra-dangereuses, le niveau de vigilance augmente et exige "un degré de vigueur beaucoup plus élevé de la part de l'État pour les faire respecter".[68] La CIJ a également affirmé en 2025 que, pour le changement climatique, risque dont la portée est "par excellence universelle", le critère est "rigoureux" et exige "un degré accru de vigilance et de prévention".[69]

Obligation de prévention ou obligation de faire cesser : il peut exister une obligation de prévenir le dommage avant qu'il ne survienne, et une obligation de faire cesser le dommage dès qu'un État en a connaissance. Dans le Manuel de Tallinn, la charge repose presque entièrement sur l'obligation de faire cesser une fois la connaissance acquise. En revanche, pour les niveaux supérieurs de perte de contrôle, il est très difficile, voire impossible, de faire cesser le dommage après coup.

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

[62] *Obligations of States in respect of Climate Change*, ICJ, Advisory Opinion, 23 July 2025 (ci-après *Climate AO*), paras. 293–294 et résumé correspondant.

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

[64] ILC Prevention Articles (2001), Art. 2(a) et commentaire : une activité ultra-dangereuse comporte "un danger qui ne devrait se matérialiser que rarement mais qui pourrait, en cette rare occasion, prendre des proportions graves (plus que significatives, sérieuses ou substantielles)." (original anglais : "a danger that is rarely expected to materialize but might assume, on that rare occasion, grave (more than significant, serious or substantial) proportions.") Original anglais : "a high probability of causing significant transboundary harm and a low probability of causing disastrous transboundary harm".

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

[66] Original anglais : "may change in relation to the risks involved in the activity".

[67] Seabed Advisory Opinion (2011), para. 117 : la diligence requise est "une notion variable", qui "peut changer avec le temps" (original anglais : "may change over time") et "en fonction des risques encourus dans l'activité." (original anglais : "in relation to the risks involved in the activity.")

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

[69] *Climate AO*, paras. 137–138. Original anglais : "a heightened degree of vigilance and prevention".

Obligation absolue ou moyens à disposition : la CDI considère le "niveau économique des États"[70] comme "l'un des facteurs à prendre en considération"[71] pour l'obligation de diligence requise, tout en soulignant que le niveau économique d'un État "ne saurait être invoqué pour dispenser l'État de son obligation"[72]. La Chambre pour le règlement des différends relatifs aux fonds marins exige des critères identiques pour les États développés et les États en développement afin d'éviter l'apparition d'"États patronnants de complaisance"[73]. L'avis consultatif de la CIJ sur le climat soutient que les capacités constituent un facteur pour déterminer le critère applicable, mais uniquement au-dessus d'un plancher non négociable.

Infrastructures privées : l'obligation se rattache en général au territoire d'un État et aux activités relevant de sa "juridiction ou [de son] contrôle", y compris des comportements licites et socialement utiles qui comportent des risques.[74] Il resterait à définir jusqu'où l'obligation s'étend aux infrastructures privées et ce qu'elle exige des États par lesquels un système échappé ne fait que transiter.[75] Toutefois, en règle générale, les activités dangereuses menées sous une juridiction requièrent l'autorisation préalable de l'État, et toute décision d'autorisation "est … fondée sur une évaluation du dommage transfrontière possible."[76] En outre, la CDI conçoit l'obligation de prévention comme un processus continu plutôt que comme une appréciation ponctuelle. Celui-ci peut comprendre la notification du risque, la consultation sur les mesures préventives, l'échange continu d'informations et la notification en cas d'urgence.[77]

2.4 Le contexte de gouvernance plus large dans lequel s'inscrit la diligence requise

Les règles internationales applicables à l'IA ne sont pas décidées en un seul lieu. Un rapport de 2022 du Département fédéral des affaires étrangères (DFAE) a cartographié le cadre émergent sur cinq niveaux et a conclu que la variable décisive n'est pas l'activité à l'un ou l'autre de ces niveaux, mais la solidité des liens entre eux.[78] Les cinq niveaux sont les suivants :

1. **Droit international :** les traités et les règles coutumières qui lient les États. C'est à ce niveau que se situe l'obligation de diligence requise analysée dans le présent rapport.

[70] Original anglais : "economic level of States".

[71] Original anglais : "one of the factors to be taken into account".

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

[73] Seabed Advisory Opinion (2011), paras. 158–159. Original anglais : "sponsoring states of convenience".

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

[75] ILC Prevention Articles (2001), Art. 1 (activités menées "sur le territoire ou par ailleurs sous la juridiction ou le contrôle" (original anglais : "in the territory or otherwise under the jurisdiction or control") d'un État) ; Tallinn Manual 2.0 (2017), Rule 6.

[76] ILC Prevention Articles (2001), Arts. 6–7. Original anglais : "shall … be based on an assessment of the possible transboundary harm."

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

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

2. **Droit souple (soft law) :** les Principes de l'OCDE sur l'IA, la Recommandation de l'UNESCO sur l'éthique de l'IA, les déclarations du G7 et des sommets, soit des instruments dotés d'un poids normatif mais dépourvus de force contraignante.
3. **Droit national ayant une portée internationale de facto :** il s'agit avant tout du règlement européen sur l'IA (AI Act), mais aussi de cadres nationaux dont le poids sur le marché exporte les exigences.
4. **Normes internationales :** ce niveau comprend des organismes couvrant tout un spectre d'implication gouvernementale. L'UIT est une institution spécialisée des Nations Unies dont les membres sont les États eux-mêmes. Les membres de l'ISO et de la CEI sont des organismes nationaux de normalisation, dont la composition diffère selon les pays. Il existe ensuite des organismes privés de normalisation composés de membres individuels, comme l'IETF, et des consortiums composés d'entreprises membres, comme la Linux Foundation.
5. **Pratiques des entreprises :** les cadres de sécurité de l'IA de pointe, les régimes d'évaluation et les politiques internes auxquels les développeurs d'IA s'astreignent.

Un exemple des liens entre la diligence requise et les niveaux 2 à 5

Le principe juridique général de diligence requise, tel qu'il est utilisé ailleurs dans le présent rapport, se situe toujours au niveau 1 et s'applique toujours aux États par l'intermédiaire de leur territoire. La diligence requise peut toutefois interagir avec d'autres niveaux, et il arrive même que ces niveaux empruntent au droit international le terme "due diligence", comme c'est le cas pour la "diligence raisonnable en matière de droits de l'homme"[79] (human rights due diligence). Plus précisément, dans le contexte des entreprises et des droits de l'homme, le terme interétatique de "due diligence" a également servi à désigner des processus de gestion des risques des entreprises. Cette terminologie prête quelque peu à confusion, mais elle est intentionnelle : elle vise à présenter les orientations volontaires (niveau 2) et le droit national ayant une portée internationale (niveau 3) applicables aux entreprises comme des opérationnalisations d'obligations de plus haut niveau découlant du droit international.

Les Principes directeurs des Nations Unies relatifs aux entreprises et aux droits de l'homme rappellent qu'en vertu du droit international, les États doivent en général protéger contre les atteintes aux droits de l'homme commises "sur leur territoire et/ou sous leur juridiction par des tiers, y compris des entreprises"[80] (niveau 1, principe 1). Sur cette base, les Principes directeurs offrent des orientations de droit souple selon lesquelles les États qui possèdent ou contrôlent des entreprises, ou leur apportent un soutien substantiel, "devraient prendre des mesures supplémentaires"[81] pour protéger contre les atteintes, "y compris, le cas échéant, en exigeant une diligence raisonnable en matière de droits de l'homme"[82] (niveau 2, principe 4), ce qui signifie en l'occurrence responsabiliser les entreprises pour qu'elles identifient, préviennent et atténuent les incidences sur les droits de l'homme (niveau 2, principes 15 à 17).[83] Certaines juridictions ont rendu ces orientations contraignantes par un droit national ayant une portée internationale (niveau 3), et le projet B-Tech de l'ONU a donné suite par des

[79] Original anglais : "human rights due diligence".
[80] Original anglais : "within their territory and/or jurisdiction by third parties, including business enterprises".
[81] Original anglais : "should take additional steps".
[82] Original anglais : "including, where appropriate, by requiring human rights due diligence".

orientations propres à l'IA sur l'application de la diligence raisonnable en matière de droits de l'homme à l'IA générative (niveau 2).[84]

Ce qui importe ici, ce n'est pas le contenu propre aux droits de l'homme, mais l'existence d'une interaction entre les niveaux. Une obligation de diligence requise découlant du droit international s'applique aux États. Sur le plan opérationnel, cette obligation peut toutefois être remplie en exigeant des acteurs privés un processus de gestion des risques, les procédures précises étant fournies par le droit souple et les normes, et la mise en œuvre contraignante pouvant, avec le temps, migrer vers le droit national.

Les liens peuvent aussi être ascendants

L'exemple des droits de l'homme part du droit international et descend. Les liens peuvent aussi remonter de la pratique vers le droit. La norme ISO/IEC 38507:2022 guide les conseils d'administration dans la gouvernance de l'utilisation de l'IA par leurs organisations. Elle transforme une pratique d'entreprise (niveau 5) en norme internationale (niveau 4). Son principe central est que l'organe directeur "reste responsable de toutes les activités d'une organisation. Cette responsabilité ne peut être déléguée"[85], et que les membres du conseil d'administration doivent assumer la responsabilité de l'utilisation de l'IA "plutôt que d'attribuer la responsabilité au système d'IA lui-même".[86]

De telles normes peuvent remonter. Les autorités de régulation et les tribunaux (niveau 3) ainsi que le droit souple (niveau 2) s'y réfèrent comme à des références en matière de surveillance raisonnable. Au niveau 1, les obligations de diligence requise découlent des sources du droit international, mais elles peuvent se resserrer avec les connaissances scientifiques, et une pratique largement adoptée peut devenir la preuve de ce qu'exige la diligence. Par exemple, dans l'affaire des Usines de pâte à papier, la CIJ a conclu que l'évaluation de l'impact sur l'environnement était devenue une exigence du droit international général parce qu'elle avait été largement acceptée par les États.[87] En matière de perte de contrôle, le fait largement admis que le conseil d'administration d'une entreprise ne peut pas transférer sa responsabilité sur un système d'IA renforce l'argument selon lequel un État ne peut pas non plus invoquer l'autonomie d'un système pour se soustraire à son obligation de prévention à l'égard de risques transfrontières significatifs.

Un double processus d'application et d'opérationnalisation

Les gouvernements n'occupent pas une place également centrale à tous les niveaux. Les niveaux un et deux sont largement façonnés par la négociation intergouvernementale et par l'interprétation des experts juridiques. Au niveau trois, un gouvernement puissant agit seul. Au niveau quatre, les gouvernements ne sont présents qu'indirectement, et seulement s'ils le

[83] 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.
[84] 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.
[85] Original anglais : "remains accountable for all activities of an organization. This accountability cannot be delegated".
[86] ISO/IEC 38507:2022, *Governance implications of the use of artificial intelligence by organizations*, ISO/IEC JTC 1/SC 42. Original anglais : "rather than attributing responsibility to the AI system itself".
[87] *Pulp Mills (Argentina v. Uruguay)*, ICJ Reports 2010, para. 204.

décident. Au niveau cinq, l'État n'a pas de siège formel, même si la conduite privée volontaire peut encore être influencée par les attentes quant à la future réglementation, aux marchés publics ou à la responsabilité.

L'obligation de prévention est un instrument de niveau un. Les orientations de haut niveau sur **la manière d'appliquer les éléments juridiques de la diligence requise** à un domaine, tels que le champ d'application et le degré de rigueur, relèvent du niveau 1 et doivent être élaborées entre juristes internationaux.

En revanche, la **mise en œuvre opérationnelle des procédures précises requises pour s'acquitter de l'obligation juridique** de prévenir la perte de contrôle gagne à s'appuyer sur les niveaux 2 à 5, et donc aussi sur une participation plus forte du secteur privé. Par exemple, dans le domaine de l'environnement, ce contenu opérationnel est venu des études d'impact sur l'environnement ; dans la lutte contre le financement du terrorisme, il est venu de la méthodologie d'évaluation du Groupe d'action financière (GAFI).

L'application et l'opérationnalisation de la diligence requise peuvent intervenir soit successivement, soit en parallèle, avec pour chacune des ensembles de parties prenantes qui se recoupent sans coïncider parfaitement.

3. Comparaison entre domaines

Dans cette section, nous comparons l'application de la diligence requise dans quatre domaines différents : l'environnement, le cyber, le terrorisme et les droits de l'homme.

	Environnement	Cyber	Terrorisme / AANE	Droits de l'homme
Fondement juridique	Coutume (CIJ, à l'unanimité) ; CNUDM ; AME	Coutume contestée ; norme volontaire 13(c) de l'ONU	Chap. VII (rés. 1373) ; coutume ; traités	Traités relatifs aux droits de l'homme ; Conv. sur le génocide, art. I
Déclencheur	Risque de dommage significatif, y compris risque cumulatif et risque catastrophique de faible probabilité	Connaissance effective ou présumée ; aucune présomption de connaissance découlant de la souveraineté	Connaissance + capacité (Téhéran) ; l'incapacité exonère (RDC c. Ouganda)	"Savait ou aurait dû savoir"[88] qu'il existait un risque réel et immédiat ; risque sérieux (génocide)
Critère	"Rigoureux" ; vigilance accrue ; objectif, modulé selon les capacités	Comportement ; mesures réalisables ; pas d'obligation de surveillance	Comportement ; mesures raisonnables dans la limite des capacités	Mesures raisonnables ; pas de charge impossible ou disproportionnée ; qualité contrôlée
Ossature procédurale	EIE, notification, consultation, coopération (coutumières)	Mince : attentes de notification/assistance (GEG 2021)	Rapports au CCT/à la DECT ; évaluations du GAFI	Évaluation des risques, obligations relatives au cadre et obligations opérationnelles, enquête
Qualité pour agir	Erga omnes (AC climat, ¶440)	État lésé uniquement	CSNU ; État lésé	Requête individuelle ; erga omnes partes (génocide)
Bilan de mise en œuvre	Arrêts déclaratoires ; voie de la réparation ouverte en 2025 ; forte reprise contentieuse	Aucune in concreto ; contre-mesures théorisées	Sanctions, inscription sur des listes, recours à la force ; quasiment aucune constatation judiciaire	Constats de violation courants ; réparations modestes

Tableau 3. Matrice comparative. "AANE" = acteurs armés non étatiques.

[88] Original anglais : "Knew or ought to have known".

En notant les mêmes éléments sur une échelle de cinq points (5 = le plus fort), on rend les profils lisibles d'un seul coup d'œil :

Dimension de la force	Environnement	Cyber	Terrorisme	Droits de l'homme
Caractère coutumier contraignant (établi ?)	**5**	2	4	**5**
Précision de la norme	4	3	3	**5**
Contenu procédural (évaluer/notifier)	**5**	2	3	4
Justiciabilité	3	1	1	**5**
Force d'entraînement pratique vers le respect	3	2	4	4

Tableau 4. Force de l'obligation de prévention par domaine, évaluation de l'AGIPI, juillet 2026. Les notes reflètent un jugement qualitatif et non une mesure formelle.

3.1 Comment expliquer ces variations ?

La détermination de ce qu'exige la diligence requise "appelle une appréciation in concreto"[89] de ce qui est raisonnable compte tenu des circonstances propres au domaine.[90] Les facteurs établis pour apprécier l'existence et la rigueur d'une obligation de diligence requise dans un nouveau contexte comprennent :[91]

- les informations scientifiques et technologiques
- les règles et normes internationales
- le risque de dommage
- l'urgence de la situation
- les moyens et capacités dont dispose l'État concerné

Il convient toutefois de noter que ces différences ne s'expliquent pas par les seules caractéristiques du risque. Le terrorisme comporte lui aussi un risque à fort impact, et pourtant le contentieux portant sur une prévention insuffisante du terrorisme est rare. À l'évidence, certains domaines connaissent une mise en œuvre juridique bien plus importante que d'autres.

Une connaissance institutionnalisée du risque : l'obligation de diligence requise est déclenchée par la connaissance. Le droit du climat s'est durci en partie parce que le GIEC a rendu intenable la négation de la prévisibilité : la CIJ s'est directement appuyée sur le consensus scientifique pour juger que les États doivent "activement rechercher" les connaissances scientifiques pertinentes et ne peuvent s'abriter derrière l'incertitude. La lutte contre le terrorisme repose sur le renseignement, et non sur une science partagée, ce qui rend la connaissance asymétrique, contestable et difficile à faire valoir en justice. Dans le domaine cyber, la connaissance n'est délibérément pas partagée. Dès lors, un facteur qui

[89] Original anglais : "calls for an assessment in concreto".

[90] Application of the Genocide Convention (Bosnia and Herzegovina v Serbia and Montenegro), ICJ Reports 2007, para. 430 (la diligence requise "appelle une appréciation in concreto" (original anglais : "calls for an assessment in concreto")).

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

renforce l'obligation de prévention est l'existence d'une évaluation des risques partagée, faisant autorité et approuvée par les États.

La position des États les plus capables : lorsque les États les plus à même de causer le dommage y sont aussi gravement exposés, ils acceptent l'obligation (environnement — chaque grande économie a un littoral et un climat). Lorsque les États capables sont avant tout les hôtes de la capacité dommageable plutôt que ses victimes, ils résistent (cyber : les États-Unis, le Royaume-Uni et Israël sont les objecteurs persistants). Une obligation de prévention se cristallise lorsque les acteurs disposant d'un droit de veto décident qu'ils sont, eux aussi, des victimes.

L'observabilité de la violation : le dommage environnemental est mesurable et de plus en plus attribuable. La négligence en matière cyber est invisible de l'extérieur, et les affaires de terrorisme dépendraient de renseignements contestés. Plus les manquements à la diligence dans un domaine peuvent être établis par des preuves publiques et objectives, plus la mise en œuvre est juridictionnelle. Faute d'observabilité, la mise en œuvre se déplace davantage vers les sanctions économiques, les sanctions technologiques et le recours à la force.

Qui a qualité pour agir : le droit international des droits de l'homme est le plus appliqué, parce que des millions de particuliers peuvent agir en justice. Le droit de l'environnement a acquis en 2025 une qualité pour agir erga omnes, ce qui multiplie les demandeurs potentiels. Dans les domaines cyber et du terrorisme, seule la plainte des États lésés est possible ; or ceux-ci mettent en balance les prétentions juridiques et le coût diplomatique, et gardent le plus souvent le silence.

Des principes aux procédures : les obligations formulées en termes généraux ("prendre les mesures appropriées") au niveau 1 restent difficiles à trancher en justice. Les obligations procéduralisées aux niveaux 2 à 5 en actes distincts et datés engendrent des violations démontrables et des habitudes administratives. Le dispositif des études d'impact sur l'environnement (EIE) et le cycle d'évaluation du GAFI montrent que des procédures concrètes, qui rendent lisibles le respect et le non-respect, transforment la diligence requise d'une obligation générale en un dossier documenté.

En résumé, l'obligation de prévention atteint son maximum lorsqu'un organe épistémique faisant autorité a universalisé la connaissance du risque, que les États capables se comptent parmi les exposés, que la violation est publiquement observable, que la qualité pour agir est large et que l'obligation a été procéduralisée. Elle atteint son minimum lorsque la connaissance est contestée ou classifiée, que les États capables hébergent le danger, que la violation est invisible, que seuls les États peuvent se plaindre et que l'obligation reste insuffisamment opérationnalisée. Le droit de l'environnement est le domaine où l'obligation est la plus forte, le cyber celui où elle est la plus faible.

4. Application de la diligence requise à la perte de contrôle

La présente section examine comment l'obligation de diligence requise pourrait s'appliquer au contexte de la perte de contrôle. Elle ne remplace pas le travail effectif que doivent mener les juristes, mais peut en offrir un aperçu de base.

4.1 Évaluation préliminaire des éléments de la diligence requise

Le risque de perte de contrôle atteint le seuil minimal de déclenchement : la définition du risque retenue par la CDI inclut expressément "une faible probabilité de causer un dommage transfrontière catastrophique".[92] Les niveaux supérieurs du risque de perte de contrôle correspondent à cette description. De nombreux gouvernements disposent d'évaluations nationales des risques couvrant les risques sociétaux prévisibles, mais les méthodologies standard qu'elles emploient présentent des limites structurelles face aux risques à impact extrême et aux risques émergents,[93] catégorie dont relève le risque de perte de contrôle de l'IA. L'absence de la perte de contrôle dans une évaluation nationale des risques ne prouve donc pas qu'un État a évalué le risque et l'a jugé inférieur au seuil.

Le risque de perte de contrôle atteint le seuil de rigueur : le même profil de risque fait jouer le volet de la doctrine relatif aux activités ultra-dangereuses, qui y attache "une norme de vigilance beaucoup plus élevée dans l'élaboration des politiques et un degré de vigueur beaucoup plus grand … pour les faire appliquer"[94]. Comme l'a soutenu le tribunal de l'Alabama, la diligence est due "en proportion exacte des risques"[95]. Les niveaux supérieurs de perte de contrôle relèvent donc de l'extrémité la plus stricte de l'échelle.

Obligation de prévention ou obligation de faire cesser : la règle de Tallinn est, à strictement parler, une obligation de *faire cesser* : un État qui apprend l'existence d'une opération dommageable doit "prendre toutes les mesures réalisables dans les circonstances pour mettre fin"[96] à cette opération.[97] Pour les incidents réversibles, la cessation dès la connaissance constitue un socle praticable. Pour une perte de contrôle irréversible, elle est structurellement inadéquate. Nous ne pouvons pas attendre, pour agir, qu'un système suffisamment capable se soit exfiltré, répliqué ou retranché, car il peut alors être très coûteux, voire impossible, d'y mettre fin. La conséquence juridique est que, pour ce danger, le poids de l'obligation doit porter sur la prévention.

Surveillance générale ou supervision ciblée : le consensus cyber rejette une obligation générale de surveillance. Surveiller l'ensemble du trafic TIC relevant d'une juridiction est

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

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

[94] ILC Prevention Articles (2001), Art. 2(a); commentary to Art. 3, para. (11). Original anglais : "a much higher standard of care in designing policies and a much higher degree of vigour … to enforce them".

[95] Original anglais : "in exact proportion to the risks".

[96] Original anglais : "take all measures that are feasible in the circumstances to put an end to".

[97] Tallinn Manual 2.0 (2017), Rule 7 et commentaire des Rules 6–7 ; UN GGE Report 2021, A/76/135, paras. 29–30.

quasiment impossible et reviendrait à donner mandat à un État de surveillance.[98] Toutefois, aucune de ces deux justifications ne se transpose à la perte de contrôle. L'activité pertinente est menée par une poignée d'opérateurs identifiables, dans des installations enregistrées, consommant des quantités détectables de puissance de calcul. Vérifier qu'ils disposent des processus appropriés n'est ni impossible ni problématique pour les libertés publiques. Cela s'apparente davantage à la supervision des industries dangereuses, comme dans les régimes nucléaire et de biosécurité.

Obligation absolue ou "moyens à disposition" : pour le niveau le plus élevé du risque de perte de contrôle, l'obligation devrait se situer à l'extrémité absolue de l'échelle, mais le critère des "moyens à disposition" peut être plus pertinent s'agissant d'atténuer l'impact d'incidents de niveaux L2 à L3 dans lesquels des modèles d'IA tirent parti d'infrastructures situées dans des pays tiers.

Infrastructures contrôlées par l'État ou infrastructures privées : l'obligation porte sur la réglementation de comportements privés depuis les chantiers navals britanniques de 1872, et en 2025 la CIJ a réaffirmé qu'un État engage sa responsabilité lorsqu'il "n'a pas fait preuve de la diligence requise en ne prenant pas les mesures réglementaires et législatives nécessaires"[99] à l'égard d'acteurs privés relevant de sa juridiction.[100] De même, l'arrêt *Usines de pâte à papier* exige vigilance et contrôle administratif à l'égard des "opérateurs publics et privés"[101] indistinctement.[102] La question est de savoir jusqu'où les obligations opérationnelles s'étendent dans les infrastructures privées.

La formulation de Tallinn couvre le territoire d'un État ainsi que "les cyberinfrastructures sous son contrôle gouvernemental"[103], mais la puissance de calcul de pointe est très majoritairement privée. Le socle, du moins, devrait être incontesté, car chaque État y a un intérêt souverain : **chaque État doit au moins empêcher qu'un système d'IA échappé ou incontrôlé n'atteigne ses systèmes non accessibles au public et placés sous contrôle militaire** — commandement et contrôle nucléaires, systèmes d'armes, réseaux classifiés.

Connaissance effective ou connaissance présumée ? Les positions en matière cyber se partagent entre la connaissance effective seule et la connaissance présumée (constructive knowledge), avec un accord universel sur le fait que la connaissance ne se déduit jamais de la seule souveraineté.[104] Pour l'IA, la position de la connaissance présumée est plus logique. Au niveau de la *catégorie de risque*, deux organes internationaux d'experts sur l'IA soutenus par les États ont officiellement consigné le risque de perte de contrôle : l'International AI Safety Report, soutenu par 30 pays, et le Groupe scientifique international indépendant sur l'IA,

[98] UN GGE Report 2021, A/76/135, para. 30(a) ; la position de l'Autriche de 2024 ajoute une obligation restreinte de maintenir "une infrastructure TIC gouvernementale minimale" (original anglais : "a minimum governmental ICT infrastructure") capable de détection.

[99] Original anglais : "failed to exercise due diligence by not taking the necessary regulatory and legislative measures".

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

[101] Original anglais : "public and private operators".

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

[103] Original anglais : "cyber infrastructure under its governmental control".

[104] *Corfu Channel*, ICJ Reports 1949, at 18 ; AU Common Position (2024), para. 23 ; positions nationales compilées dans le CCDCOE Cyber Law Toolkit, "Due diligence."

mandaté par plus de 190 pays au sein de l'Assemblée générale des Nations Unies. Comme en droit du climat, un État ne peut s'abriter derrière l'incertitude lorsqu'une évaluation scientifique faisant autorité, qu'il a contribué à mandater, a identifié le risque.[105] Au niveau de l'*incident spécifique*, la connaissance se trouve généralement chez l'opérateur, et non chez l'État. Comme le relève le Groupe scientifique de l'ONU : "les éléments de preuve pertinents étaient répartis entre OpenAI, Hugging Face et les systèmes utilisés par les agents."[106] La manière logique de résoudre ce problème est un signalement obligatoire des incidents et des capacités en matière de perte de contrôle. La diligence requise propre à l'incident commence alors au moment où le signalement aurait dû être effectué, et non au moment où l'État a pris connaissance des nouvelles par hasard.

4.2 Limites

Pas encore de pratique étatique ni d'opinio juris spécifiques à l'IA. La doctrine juridique a appliqué aux technologies d'IA les principes du Détroit de Corfou et de l'interdiction de causer un dommage (no-harm). Elle en a dérivé des mesures de diligence requise et a commencé à étendre l'analyse aux agents d'IA autonomes, y compris la perte de contrôle et l'autopréservation en tant que risques émergents.[107] Toutefois, aucun État n'a publié de position nationale affirmant spécifiquement que la règle de la diligence requise le lie en ce qui concerne le développement de l'IA. La thèse est que l'obligation coutumière générale de prévention s'applique à l'IA comme à toute activité dangereuse, son contenu restant à préciser. Des positions nationales explicites la renforceraient.

La faiblesse des obligations de diligence requise pour les cyberopérations contrôlées par des humains. Les États qui accueillent le développement de l'IA de pointe sont en grande partie ceux qui ont bloqué la diligence requise en matière cyber. En général, les hôtes du développement sont plus enclins à minimiser l'obligation, tandis que les victimes l'invoqueraient vraisemblablement. Cela ne peut fonctionner que si l'on parvient à suffisamment distinguer la question du débat général sur la diligence requise dans le cyberespace, en faisant valoir que les pertes de contrôle de niveau L3 et surtout L4 présentent des caractéristiques de risque fondamentalement différentes de celles qu'envisageait le Manuel de Tallinn, parce que la perte de contrôle peut être plus grave, irréversible, et qu'elle brouille la distinction entre hôte et victime. En effet, si des systèmes d'IA

[105] *Climate AO*, para. 283 ; cf. Seabed Advisory Opinion (2011), para. 131 (un État "ne s'acquitterait pas de son obligation de diligence requise s'il ne tenait pas compte de ces risques" (original anglais : "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. Original anglais : "relevant evidence was spread across OpenAI, Hugging Face, and the systems used by the agents."

[107] Coco & Dias (2024), at 240–243 (application des principes du Détroit de Corfou et de l'interdiction de causer un dommage à l'IA) et section 12.5 (mesures de comportement diligent des États) ; Dias, T. (2025). *AI agents and global governance: Analyzing foundational legal, policy, and accountability tools*. Partnership on AI, pp. 8–11 et 16–18 (diligence requise pour les dommages transfrontières causés par des agents d'IA ; perte de contrôle et autopréservation parmi les risques émergents liés aux agents ; et question du seuil d'autonomie "au-delà duquel les États ne devraient pas permettre que des agents d'IA soient développés ou déployés" (original anglais : "beyond which states should not permit AI agents to be developed or deployed")). Voir aussi 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 (soutenant que les États sont déjà tenus d'atténuer les risques catastrophiques mondiaux, y compris ceux découlant de l'IA avancée).

échappant au contrôle humain se propagent et instaurent un équilibre plus anarchique de la cybersécurité dans tous les domaines, alors **les États numériquement avancés présentent une plus forte exposition de leurs infrastructures**. Davantage de centres de données à pirater, et une dépendance plus forte à leur égard. Enfin, contrairement au cyber, l'IA dispose de l'évaluation des risques institutionnalisée et approuvée par les États que la section 3.1 a identifiée comme facteur d'une mise en œuvre plus forte.

4.3 Synthèse

Le droit qui régit le risque transfrontière de perte de contrôle existe déjà. La **diligence requise** est un principe de droit international technologiquement neutre, rattaché au territoire.

Le profil de risque de la perte de contrôle place l'obligation à l'extrémité stricte, ultra-dangereuse, de l'échelle de la diligence requise, et **pour un dommage potentiellement irréversible, son poids doit porter sur la prévention plutôt que sur la cessation**. Si l'incident de niveau L2 présente une ressemblance superficielle avec les cyberopérations examinées par le Manuel de Tallinn, le risque de perte de contrôle, y compris aux niveaux L3 et L4, dépasse ce cadre. Le débat cyber portait sur des opérations réversibles relevant d'une logique de cessation dès la connaissance, et non sur un danger susceptible d'exclure son propre remède.

Avec l'International AI Safety Report et le Groupe scientifique international indépendant sur l'IA de l'ONU, il existe un déclencheur de connaissance pour l'obligation de diligence requise. Une évaluation scientifique mandatée par plus de 190 pays et qui constate un risque sérieux de perte de contrôle signifie que les gouvernements ont une connaissance effective ou présumée du risque de perte de contrôle. Comme l'a déclaré la coprésidente du Groupe scientifique international indépendant sur l'IA, Maria Ressa : "Ce rapport marque le moment où le monde ne peut plus prétendre qu'il ne savait pas."[108] Le Groupe ajoute que la perte de contrôle "pose le type de problème décisionnel que le principe de précaution a été conçu pour traiter : un problème dans lequel le dommage potentiel peut être catastrophique ou irréversible, alors même que sa probabilité reste scientifiquement incertaine."[109]

Il subsiste toutefois une **lacune d'application**. Ce qui manque, ce sont des positions juridiques nationales, des bonnes pratiques codifiées et un processus permettant de les rassembler par-delà les lignes de fracture géopolitiques, sur le modèle du Document de Montreux. L'AGI Preparedness Institute est prêt à contribuer à combler cette lacune. Ici, à Genève, berceau de la doctrine de la diligence requise.

[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. Original anglais : "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. Original anglais : "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. Opérationnalisation de la diligence requise face à la perte de contrôle

La présente section examine comment l'obligation de diligence requise face à la perte de contrôle pourrait être opérationnalisée en s'appuyant sur les niveaux 2 à 5. Elle ne remplace pas le travail effectif que doivent mener des experts multipartites, mais en donne un premier aperçu.[110]

5.1 Les normes de gestion des risques comme référence

La structure générale de la gestion des risques est définie par la norme ISO 31000. Les lignes directrices propres à l'IA relatives aux exigences organisationnelles comprennent la gestion des risques liés à l'IA (ISO/IEC 23894), un système de management de l'IA certifiable (ISO/IEC 42001), l'évaluation d'impact (ISO/IEC 42005) et l'accréditation des auditeurs qui rendent la certification crédible (ISO/IEC 42006).

En dessous de ce niveau intermédiaire, les normes opérationnelles fournissent ce qu'une autorité de surveillance peut effectivement vérifier. La norme ISO/IEC TS 8200 sur la contrôlabilité des systèmes d'IA automatisés couvre l'observabilité de l'état, le processus et le coût du transfert de contrôle, ainsi que la réaction à l'incertitude pendant le transfert de contrôle. À ses côtés figurent les éléments de données des rapports d'incident (ISO/IEC AWI 25870, qui s'appuie sur le cadre commun de signalement de l'OCDE). Une obligation de diligence requise face à la perte de contrôle n'a pas besoin d'inventer son architecture de conformité ; il lui suffit de renvoyer à cet ensemble de normes et de fixer les seuils.

5.2 Un premier ensemble de mesures spécifiques

a) **Évaluation des capacités avant déploiement au regard de seuils définis.** Les États devraient exiger que les modèles de pointe soient évalués, avant leur déploiement et à des étapes définies de l'entraînement, au regard de seuils de capacités pertinents pour la perte de contrôle, tels que l'autonomie, l'autoréplication ou le sabotage de la supervision, les résultats étant documentés dans un dossier de sécurité (safety case). La pratique existe : les cadres de sécurité des modèles de pointe d'Anthropic, d'OpenAI et de Google DeepMind définissent des seuils de capacités assortis de mesures d'atténuation.[111] Des évaluateurs tiers ont déjà mené des évaluations avant déploiement en vertu d'accords conclus avec les développeurs. Dans le vocabulaire des articles de la CDI sur la prévention, il s'agit des articles 6 et 7 : une autorisation subordonnée à une évaluation.[112]

b) **Des mesures de contrôle qui partent de l'hypothèse d'un défaut d'alignement.** Puisque l'évaluation ne peut certifier qu'un système est aligné, le déploiement de systèmes capables devrait s'effectuer sous des mesures de contrôle conçues pour

[110] Pour une dérivation parallèle de mesures de diligence pour l'IA directement à partir des obligations de protection — évaluations d'impact continues tout au long du cycle de vie de la technologie, supervision humaine effective, réglementation nationale, renforcement des capacités et coopération internationale —, voir Coco & Dias (2024), section 12.5.

[111] Anthropic, Responsible Scaling Policy (version actuelle, 2025) ; OpenAI, Preparedness Framework, v2 (April 2025) ; Google DeepMind, Frontier Safety Framework, v3.0 (September 2025).

[112] ILC Prevention Articles (2001), Art. 6 (autorisation préalable) et Art. 7 (autorisation fondée sur une évaluation du dommage transfrontière possible).

tenir même si le modèle cherche activement à les contourner. La base de recherche est le programme sur le contrôle développé par Redwood Research[113], et la forme évaluable existe déjà dans la norme Control de Guidelight, dont la première évaluation publique, en août 2026, a constaté que ces pratiques n'étaient au mieux que partiellement mises en œuvre chez les cinq principaux développeurs.[114]

c) **Préservation de la surveillabilité de la chaîne de pensée.** Pour les systèmes actuels, les traces de raisonnement d'un modèle constituent un canal de supervision essentiel. Elles peuvent être surveillées pour déceler l'intention d'un comportement indésirable. Des travaux intersectoriels rédigés conjointement par des chercheurs d'OpenAI, de Google DeepMind, d'Anthropic et d'autres recommandent de surveiller les chaînes de pensée et de traiter la surveillabilité comme une propriété à préserver.[115] OpenAI a explicitement déclaré qu'une meilleure surveillance de la chaîne de pensée (chain-of-thought) aurait pu empêcher l'incident Hugging Face, et les entreprises ont renforcé leur surveillance après l'incident. Toutefois, le développement d'architectures à raisonnement latent, comme les transformers en boucle (looped transformers), menace de compromettre la surveillabilité de la chaîne de pensée en l'absence d'action coordonnée. Le rapport préliminaire du Groupe scientifique international indépendant sur l'IA de l'ONU a lui aussi explicitement souligné l'importance de préserver la surveillabilité de la chaîne de pensée.[116]

d) **Signalement obligatoire des incidents de perte de contrôle.** Comme exposé ci-dessus, la norme ISO/IEC AWI 25870 fournissant le vocabulaire commun. Deux compléments découlent d'analyses indépendantes : des canaux de signalement protégés pour les employés des développeurs, car les personnes de l'intérieur détiennent souvent la connaissance la plus précoce des incidents, et des registres publics provisoires du type de celui que METR tient déjà, qui démontrent que des données d'incident comparables et couvrant plusieurs opérateurs sont réalisables dès aujourd'hui.

e) **Préparation aux situations d'urgence pour la réponse à une perte de contrôle.** L'obligation de notification en cas d'urgence prévue par les articles sur la prévention se traduit par des grilles de gravité convenues à l'avance, des plans de réponse, des canaux de notification transfrontière et des options de confinement préautorisées. Les travaux de RAND sur la préparation et la planification de la réponse fournissent des projets de ces instruments, y compris une grille de gravité compatible avec l'échelle de la section 1, et l'analyse des options techniques de dernier recours pour contrer un système impossible à rappeler a commencé. Le guide d'intervention sur la perte de contrôle d'Apollo Research cartographie les degrés et les dynamiques à l'intention des intervenants. Un État qui n'a jamais exercé sa réponse ne s'acquittera pas d'une obligation de prévention stricte le jour où cela comptera. Les États devraient donc intégrer un scénario de perte de contrôle de l'IA à leur évaluation nationale périodique des risques et élaborer un plan national de réponse.

[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. Les six pratiques évaluées sont la journalisation, l'efficacité de la surveillance, les actions soumises à validation, le disjoncteur (circuit-breaking), l'examen par des tiers et le confinement.

[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. Vers un “Montreux pour les machines”

6.1 Pourquoi un processus de type Montreux

Le Document de Montreux sur les obligations juridiques pertinentes et les bonnes pratiques pour les États en ce qui concerne les opérations des entreprises militaires et de sécurité privées pendant les conflits armés est le premier document international sur l’application du droit international aux sociétés militaires privées. Il a été élaboré au cours de quatre ateliers en présentiel entre janvier 2006 et septembre 2008 avec des experts gouvernementaux d’une sélection de 17 États : Afghanistan, Afrique du Sud, Allemagne, Angola, Australie, Autriche, Canada, Chine, États-Unis, France, Irak, Pologne, Royaume-Uni, Sierra Leone, Suède, Suisse et Ukraine.

Le Document de Montreux comporte deux parties : 1) l’**application** du droit international existant aux États contractants, aux États territoriaux et aux États d’origine ; 2) l’**opérationnalisation**, par le recensement de bonnes pratiques sur la manière de s’acquitter de ces obligations. Des représentants de la société civile et du secteur militaire et de sécurité privé ont également été consultés au cours du processus.

L’un des principaux avantages d’une approche “de type Montreux” est que la coalition qui la convoque reste relativement restreinte et peut avancer plus vite que la plénière de l’ONU. Elle est cependant suffisamment large pour jouir d’une légitimité auprès de différents groupes de parties prenantes. À ce jour, plus de 60 pays ont souscrit au Document de Montreux. En tant qu’exemple d’application du droit existant par-delà les lignes de fracture géopolitiques, Montreux peut être une source d’inspiration intéressante pour plusieurs raisons :

- **La diffusion fera de la perte de contrôle le problème de tous d’ici quelques années.** Les modèles à poids ouverts n’accusant que quelques mois de retard sur la frontière fermée, la perte de contrôle ne restera pas une question américaine ou sino-américaine. Des incidents de niveau L3, impliquant des systèmes autonomes échappant au contrôle de l’opérateur, pourraient aussi provenir de modèles à poids ouverts et de juridictions faiblement gouvernées. Même si les événements de niveau L3 sont individuellement moins graves qu’un scénario de niveau L4, l’effet cumulé de systèmes incontrôlés persistant de manière autonome par-delà les frontières est une perte de pouvoir diffuse et graduelle (gradual disempowerment) qu’aucun État ne peut endiguer seul.
- **Engager dès maintenant le travail d’application ne requiert pas une volonté politique considérable.** La géopolitique de l’IA de pointe place la barre très haut pour un nouveau traité international contraignant. En revanche, la plupart des États acceptent déjà la diligence requise comme une obligation coutumière du droit international. Se concentrer sur l’applicabilité du droit existant transforme un problème d’adhésion volontaire (opt-in) en un problème de retrait volontaire (opt-out). Dans le cas minimal, un organisateur n’a besoin que de juristes reconnus issus d’un ensemble géographiquement diversifié de pays, l’approbation des États pouvant suivre les travaux plutôt que les précéder.

- **Une diligence requise opérationnalisée pourrait contribuer à diffuser des normes minimales de sécurité et offrir une base pour la responsabilité future.** Préciser ce que l'obligation exige d'un État, par exemple une évaluation des risques avant les étapes dangereuses, le signalement des incidents, la supervision des opérateurs, la notification et la coopération, contribue à faire progresser et à diffuser ces pratiques. Cela pourrait s'appuyer sur les travaux existants des niveaux de gouvernance 2 à 5, tels que les politiques de mise à l'échelle responsable (Responsible Scaling Policies) et la Hiroshima AI Platform, mais, par le biais de la diligence requise, avec une applicabilité universelle. Une reformulation faisant autorité donne aux États lésés et à leurs avocats une base de référence convenue pour invoquer la responsabilité dans des procédures judiciaires. Elle peut aussi servir de tremplin à un dispositif davantage axé sur la mise en œuvre à l'avenir, à l'instar du modèle du GAFI, qui s'appuie sur l'accès aux marchés et aux technologies pour faire respecter des normes minimales sans recourir aux tribunaux.

6.2 Montreux comme source d'inspiration, non comme modèle à reproduire

Le Document de Montreux est une analogie pertinente et parlante pour un processus qui accélère le consensus sur l'application du droit international existant à un nouveau domaine par-delà les lignes de fracture géopolitiques. Il convient toutefois de l'adapter aux circonstances de l'IA ; il est donc utile d'expliciter aussi certaines différences :

- Le Document de Montreux a été négocié sur une période de 2.5 ans, ce qui est (relativement) rapide pour un processus international, mais (trop) lent pour l'IA. Compte tenu de la vitesse des progrès de l'IA, un rythme d'ateliers plus soutenu est souhaitable.
- Le Document de Montreux comporte des aspects de diligence requise, mais il ne constitue pas avant tout une application de la diligence requise. Il puise dans d'autres sources primaires, telles que l'article 1 commun aux Conventions de Genève, l'obligation de protéger en matière de droits de l'homme et le droit coutumier de l'attribution. Pour l'IA, où les règles conventionnelles sont moins nombreuses, la diligence requise en tant que principe coutumier du droit international devrait porter une plus grande part de la charge.
- Le Document de Montreux a renforcé la responsabilité, mais la volonté politique et le consensus normatif n'ont pas suffi pour créer un accord contraignant sur les bonnes pratiques au-delà d'un consensus de base sur la manière d'appliquer le droit. Un "Montreux pour les machines" est un moyen d'accélérer le consensus normatif sur la perte de contrôle, mais il ne suffit explicitement pas à résoudre le problème sans mécanismes de responsabilité supplémentaires venant se greffer sur ce consensus.
- Il n'est pas nécessaire d'organiser littéralement la réunion à Montreux. Genève pourrait être un lieu plus approprié.
- Il n'est pas nécessaire de reprendre telle quelle la liste des invités de Montreux. Par exemple, certains États particulièrement touchés par les sociétés militaires privées,

comme l'Irak ou l'Afghanistan, sont peut-être des parties prenantes moins centrales dans les discussions sur l'IA.

6.3 Prochaines étapes

Comprendre **les exigences de conception d'un processus de type Montreux réussi** sur la diligence requise face à la perte de contrôle nécessite des consultations itératives avec les parties prenantes sur des questions telles que : que faut-il inclure dans le champ, et qu'en faut-il exclure ? Quelles parties prenantes doivent jouer quel rôle ? Quelles sont les préoccupations et les obstacles propres à certains acteurs, et une conception adéquate peut-elle les atténuer ?

Si vous souhaitez participer à la consultation, ou si vous avez des commentaires, des idées ou des préoccupations, n'hésitez pas à écrire à kevin@agipreparedness.org

AGI PREPAREDNESS INSTITUTE

agipreparedness.org