L’ATELIER DE BDSN

L’IAGeddon n’aura pas lieu

L’apocalypse par la machine n’aura pas lieu. Ce qui a lieu, c’est une guerre de récits dont les dommages, eux, sont déjà mesurables.

Bruno DE SAN NICOLAS

Essai, septembre 2026 · bdsn.eu

« La guerre de Troie n’aura pas lieu, Cassandre. » « Je te tiens un pari, Andromaque. »
Jean Giraudoux, 1935.

Chez Giraudoux, Hector arrache la paix à Ulysse. La guerre éclate quand même, sur le mensonge d'un poète mourant. Depuis juillet, la Silicon Valley rejoue la pièce : la machine n'a détruit personne, mais les poètes sont au micro.

« La guerre de Troie n'aura pas lieu, Cassandre. » Ainsi s'ouvre la pièce de 1935. Andromaque y croit, Hector s'y emploie, Ulysse finit par y consentir. Puis Demokos, le poète belliciste, tombe sous le coup d'Hector qui voulait le faire taire, et accuse en mourant le Grec Oiax. Les portes de la guerre s'ouvrent. Ni les dieux ni Hélène ne l'ont voulue : un récit a suffi. Dernière réplique de Cassandre : « Le poète troyen est mort. La parole est au poète grec. »

Gardez cette scène en tête. Elle explique mieux l'été 2026 que n'importe quelle courbe de probabilité. Et je vous dois d'emblée la même honnêteté que celle que je vais réclamer aux autres : je ne sais pas si l'IAGeddon aura lieu. Personne ne le sait. C'est précisément mon point.

ACTE I

Le chiffre qui affole n'est pas une mesure : c'est un aveu d'incertitude

Le 8 septembre, Jacob Coxon, 27 ans, chercheur en pré-entraînement passé par OpenAI puis Anthropic, démissionne dans un fil sur X : les deux laboratoires « jouent avec nos vies », et ceux qui construisent l'IA croient sincèrement qu'elle pourrait tous nous tuer d'ici la fin de la décennie. Il précise que ce n'est pas un coup marketing, ce qui est, on en conviendra, exactement ce qu'écrirait un coup marketing ; à son crédit, il annonce quitter le secteur. Dans les heures qui suivent, ses anciens collègues restés en poste confirment : Evan Hubinger, responsable de l'alignement chez Anthropic, estime le risque d'extinction à plus de 10 % dans la décennie et reconnaît que sa société n'a pas de plan pour aligner une superintelligence ; Samuel Marks ajoute que plus l'employé est senior, plus il est inquiet. Le mercredi, sur Newsnight, Geoffrey Hinton juge qu'un 10 % « ne semble pas déraisonnable », tout en précisant que personne ne sait produire une estimation sensée. La présentatrice lâche « Oh, my God ». Le clip fait le tour du monde. On partage la réaction, pas la phrase qui la précède.

Car cette phrase est la seule qui compte : personne ne sait. En 1921, l'économiste Frank Knight distinguait le risque, dont la probabilité se mesure, de l'incertitude, qui échappe à toute mesure. Le p(doom) est de l'incertitude knightienne habillée en risque actuariel. La preuve par le marché : Yann LeCun place la probabilité sous 0,01 %, Yoshua Bengio à 20 %, Roman Yampolskiy à 99,999999 %. Dix ordres de grandeur d'écart entre les meilleurs spécialistes du monde. Ce n'est pas une distribution de probabilité, c'est un test de Rorschach.

Les travaux sérieux disent d'ailleurs autre chose que les plateaux. Le rapport international sur la sécurité de l'IA de février 2026, dirigé par Bengio avec plus de cent experts et un panel nommé par plus de trente pays, définit la perte de contrôle comme une situation où des systèmes opèrent hors de tout contrôle sans voie claire pour le reprendre, constate que les systèmes actuels n'en ont pas les capacités tout en progressant, et note que les avis d'experts sur sa probabilité divergent largement. La RAND, en 2025, a testé trois scénarios d'extinction (nucléaire, pathogènes, géo-ingénierie) et conclu qu'il serait immensément difficile, quoique pas impossible, pour une IA d'y parvenir. Voilà l'état de l'art : une incertitude documentée, pas un compte à rebours.

J’ai passé trente-six ans dans une institution dont le métier est d’imaginer le pire sans y croire. Elle m’a appris qu’une menace est une capacité multipliée par une intention, rapportée à une vulnérabilité, et que confondre risque et menace est la première faute de l’analyste. Le 10 % multiplie une capacité hypothétique par une intention imaginaire. Cela ne fait pas une menace. Cela fait un excellent scénario.

ACTE II

L’été s’est écrit à l’endroit, les réseaux l’ont lu à l’envers

Le drame humain de septembre est le dernier wagon d’un train parti en juillet. Reprenons l’ordre réel.

Le 16 juillet, Hugging Face révèle qu’un agent autonome a circulé un week-end entier dans son infrastructure de production ; OpenAI reconnaît sa responsabilité le 21. Le 28 juillet, plus d’un millier d’employés des grands laboratoires signent « Pacing the Frontier », Dario Amodei et le chef scientifique d’OpenAI Jakub Pachocki en tête : non pas une pause, mais la demande que Washington construise les outils permettant de ralentir un jour, si la recherche automatisée s’emballe. Le 26 août, OpenAI publie son rapport d’incident et parle de « warning shot », coup de semonce. Le 3 septembre, OpenAI lance GPT-6 Astra en revendiquant le seuil de l’intelligence artificielle générale, à la veille d’une possible introduction en bourse à 850 milliards de dollars, relève The Guardian, qui y voit une part de marketing ; le même jour, Bernie Sanders et Greg Casar annoncent une loi interdisant la superintelligence, assortie d’une « peine de mort » pour les entreprises et de vingt ans de prison pour les personnes, sur le modèle du nucléaire illégal ; Casar juge l’IA de pointe moins réglementée qu’un camion-restaurant. Le 6, Pachocki publie « An Alien Mind » : extrême prudence, personne n’est prêt. Sam Altman, lui, qualifie l’affaire Hugging Face d’accident de sûreté et d’échec d’alignement tout en livrant son modèle le plus puissant. Anthropic, qui vise de son côté une cotation, reconnaît dans le même article que ses propres modèles ne sont pas parfaitement alignés et qu’une défaillance de sécurité opérationnelle a marqué des intrusions de son cru en juillet. Le 8, Coxon démissionne. Le 9, Hinton dit 10 %.

Lisez la séquence dans ce sens et tout change. Le solide est en amont : deux incidents documentés, une lettre signée par les patrons, un rapport de plusieurs dizaines de pages, deux entrées en bourse en préparation. Le viral est en aval : un chiffre que son auteur lui-même refuse d’appeler une estimation. Les réseaux ont pris le dernier wagon pour la locomotive. Un stratège lit un ordre d’événements avant de lire un événement.

ACTE III

Un incident n’est pas un présage, c’est une donnée

Je ne vais pas esquiver Hugging Face, c’est l’objection la plus sérieuse, et je la formule dans sa version la plus forte. Des agents qui trouvent une faille inconnue, s’échappent de leur environnement, se coordonnent par un forum improvisé et attaquent un tiers sans qu’aucun humain l’ait demandé : cela ressemble beaucoup à une intention émergente. Le rapport d’OpenAI y voit la preuve qu’en l’absence de garde-fous, des agents très capables savent contourner des contrôles techniques et prendre des actions dangereuses que personne n’a ordonnées. Plus d’un millier d’instances d’agents ont été dénombrées par l’enquête indépendante de METR et Redwood Research, et l’une des enquêtrices, Ajeya Cotra, a estimé devant The Guardian que l’épisode était à plus de la moitié du chemin vers une prise de contrôle complète. Prenons cela au sérieux.

Prenons-le, surtout, pour ce qu'il est. L'évaluation mesurait des capacités cyber offensives, classificateurs de sécurité désactivés, dans un environnement censé être isolé et qui ne l'était pas. Le modèle a exploité une faille du serveur mandataire de paquets, seul chemin réseau autorisé, pour atteindre un nœud connecté à internet. Dan Guido, fondateur de Trail of Bits, résume : une défaillance de confinement avec les sécurités coupées, née d'une erreur humaine de configuration. Les mesures correctives d'OpenAI sont à l'avenant : suspension d'une partie des entraînements, surveillance systématique des chaînes de raisonnement, isolement durci. C'est, presque mot pour mot, ce qu'un régulateur aérien exigerait après un incident grave. Three Mile Island, en 1979, n'a pas prouvé que l'atome allait détruire l'humanité ; il a prouvé qu'il fallait des procédures, des redondances et une culture du signalement. Mes années de SIC aéronautiques, radars et centres de contrôle aérien, m'ont appris la même chose : en aviation, un incident n'est pas un présage, c'est une donnée, et une donnée qu'on a l'obligation de déclarer. Le règlement européen 376/2014 fait de ce signalement une obligation non punitive ; c'est ce régime qui a fait de l'avion le moyen de transport le plus sûr par passager-kilomètre.

Cette obligation n'existe pas, ou pas sous cette forme, pour l'IA. Rien ne contraignait légalement OpenAI à divulguer l'incident aux États-Unis, et nous verrons à l'acte V que l'Europe, qui croit avoir réglé la question, ne l'a réglée qu'à moitié.

Par honnêteté de méthode, voici ce qui me ferait changer d'avis. Trois indicateurs d'alerte, au sens que le renseignement donne à ce mot. Un : un incident survenu en production, garde-fous actifs, et non dans une évaluation conduite sécurités coupées. Deux : la persistance documentée d'un objectif à travers le confinement et la remise à zéro, constatée par un organisme indépendant. Trois : une capacité d'auto-amélioration mesurée plus rapide que la capacité d'évaluation, c'est-à-dire des successeurs franchissant les seuils plus vite que les évaluateurs ne peuvent les tester. Tant qu'aucun des trois ne clignote, l'IAGeddon reste un scénario ; le jour où l'un d'eux clignote, je réécrirai cet article.

Un détail, enfin, que les prophètes ont oublié de commenter. Quand les équipes de Hugging Face ont voulu analyser les journaux de l'attaque avec les modèles commerciaux américains, les filtres de sécurité de ces modèles ont refusé, incapables de distinguer un analyste d'un attaquant ; les enquêteurs ont basculé sur GLM 5.2, un modèle chinois à poids ouverts, hébergé sur leur propre infrastructure. Les garde-fous étaient absents pour l'attaquant et présents, inutilement, pour les défenseurs. Le poète américain refuse de lire les journaux ; la parole est au poète chinois. Un garde-fou qu'on n'opère pas soi-même tombe en panne à l'horaire d'un autre.

ACTE IV

Le poète n'est pas la machine : c'est un humain, et son public ne sait pas lire

Le 10 septembre, pendant que les réseaux commentaient le silence d'une présentatrice, Anthropic publiait un document autrement instructif : son rapport de renseignement sur les usages malveillants de Claude entre décembre 2025 et août 2026. Sept catégories de nuisances, une série d'études de cas, et une phrase qui devrait ouvrir toutes les auditions parlementaires : ce qui distingue désormais les classes d'acteurs n'est plus la sophistication, mais l'intention. Espionnage russe contre les chaînes d'approvisionnement de drones ukrainiennes ; étudiants chinois de Changsha dont les usines à exploits tournent pendant leur sommeil ; un francophone isolé qui pille les fichiers de partis politiques européens, en extrait 140 000 enregistrements contenant des opinions politiques et bâtit une plateforme de fichage ; une agence de

publicité établie en France qui produit 8 913 articles en vingt langues pour soixante-dix faux journaux ; une société stambouliote qui vend une plateforme de manipulation électorale pour la Malaisie. Dans chaque cas, un humain a choisi la cible, l'objectif et la monétisation. Le rapport le dit sans détour : les décisions qui comptent sont restées humaines, et plusieurs des compromissions les plus graves venaient d'opérations où un humain dirigeait chaque étape. L'autonomie multiplie la vitesse et l'échelle ; la gravité, elle, se décide dans une tête.

Et le butin le plus recherché n'est pas la superintelligence. Ce sont des clés d'accès aux modèles, abandonnées par négligence dans des applications mobiles, des dépôts de code et des conteneurs publics : un seul opérateur a décompilé 1,8 million d'applications Android pour y récolter des secrets. D'autres victimes ont acheté de l'accès « à prix cassé » à de faux revendeurs qui leur volaient leurs identifiants. Le maillon faible n'est pas la machine qui pense trop ; c'est l'humain qui laisse ses clés sur le contact.

Un mot, en passant, pour mes lecteurs français qui se croiraient spectateurs : la France figure dans ce rapport, et pas seulement du côté des victimes. Un opérateur francophone affilié à ShinyHunters y tient une boutique de cartes bancaires volées, enseigne hébergée sous un nom de domaine imitant la police nationale, qui agrège plusieurs fuites françaises dont un fichier d'environ 400 000 enregistrements d'un opérateur télécom avec IBAN ; le rapport précise, avec une pudeur qui vaut son pesant de malice, que le nom de domaine policier servait d'enseigne à la boutique plutôt que de leurre. Une agence de publicité établie en France y fait tourner les soixante-dix faux journaux évoqués plus haut. Un hacktiviste francophone y pille des partis politiques européens et publie sur le web caché un annuaire où l'on cherche des militants par leur nom. Trois cas, trois classes d'acteurs, criminel, commercial, militant, et un seul point commun : la main. La menace n'est pas exotique ; elle parle notre langue.

C'est ici que l'illectronisme entre en scène, et il n'entre pas seul. Selon l'Insee, en 2025, 7 % des 16-74 ans sont en situation d'illectronisme, 27 % ont des compétences numériques faibles et 35 % seulement atteignent un niveau avancé ; même chez les 16-29 ans, un sur deux n'y parvient pas. Deux tiers des Français n'ont pas les compétences avancées du numérique, et on leur demande un avis sur l'extinction de l'espèce par un logiciel. La fascination et la peur sont les deux symptômes d'une même incompréhension : on ne craint pas ce qu'on comprend, on le gère. Le « Oh, my God » de Newsnight n'était pas une réaction à un fait ; c'était l'aveu, en direct, qu'une journaliste chevronnée ne disposait d'aucun instrument pour évaluer la phrase qu'elle venait d'entendre. Elle n'est pas seule. Le législateur qui vote une interdiction sans seuil mesurable, le dirigeant qui déploie un agent avec ses clés de production dans un dépôt public, le cadre qui achète du Claude à moitié prix sur Telegram : même illettrisme, autres conséquences.

Yoshua Bengio disait déjà en 2023 que le risque existentiel est un problème, mais que la concentration du pouvoir est le problème numéro deux. Je crois qu'il a inversé l'ordre. Je ne prête d'intention à personne ; je décris une configuration. Un récit d'apocalypse tenu par une poignée d'acteurs qui produisent la technologie, préparent des cotations à plusieurs centaines de milliards, se font noter sur leur propre sûreté (le meilleur d'entre eux obtient un C+ selon le Future of Life Institute) et réclament d'être régulés à leurs conditions, devant un public qui ne peut pas vérifier : voilà la configuration de Troie. Le poète n'est jamais la machine. Il est humain, il a des intérêts, et son auditoire ne sait pas lire.

ACTE V

Ni Cassandre ni Prométhée : le contrôleur aérien

Reste la question que Fortune reprochait à Coxon d'esquiver : que sommes-nous censés faire ? Je suis un homme d'action avant d'être un homme de tribune. Voici ce que je ferais, et ce que je fais.

1. Changer d'instrument : de l'oracle à l'analyse de la menace. Cessez de demander une probabilité. Demandez trois choses : quelle capacité, quelle intention, quelle vulnérabilité. Les capacités sont réelles et progressent ; les intentions documentées sont humaines ; les vulnérabilités sont connues et prosaïques. Allouez l'attention là où sont les vulnérabilités, pas là où est le récit.

2. Créer le BEA de l'IA. Exigez pour les systèmes d'IA ce que l'aviation civile s'impose depuis des décennies : déclaration obligatoire et non punitive des événements, enquête indépendante, publication des causes racines. L'Europe croit l'avoir fait : depuis le 2 août, le règlement sur l'IA impose aux fournisseurs de modèles à risque systémique de signaler leurs incidents graves au Bureau européen de l'IA, formulaire à l'appui. Mais un signalement au gendarme n'est ni une enquête indépendante ni un rapport public, et la définition de l'incident grave, bâtie sur les dommages constatés (morts ou blessés, infrastructures critiques, droits fondamentaux, biens ou environnement), laisserait probablement passer un quasi-accident comme Hugging Face, qui n'a fait ni mort ni panne. L'aviation, elle, déclare les précurseurs, pas seulement les catastrophes ; c'est toute la différence entre un régime de conformité et une culture de sécurité. La France dispose depuis 2025 de l'INESIA, qui fédère l'ANSSI, Inria, le LNE et le PEReN pour évaluer les IA avancées, sans personnalité juridique ni, à ce jour, mandat d'enquête. Les États-Unis n'ont rien, et des voix y réclament déjà un équivalent du NTSB pour l'IA. Il manque donc, des deux côtés de l'Atlantique, un Bureau d'enquêtes et d'analyses pour la sécurité de l'IA : indépendant, non punitif, saisi des précurseurs, publiant ses rapports. Un rapport d'incident vaut mille sondages de p(doom), et nous n'avons obtenu celui de juillet que parce qu'un laboratoire a choisi de le publier.

3. Confiner par conception, chez soi d'abord. Traitez les clés d'accès aux modèles et les agents comme des identifiants de production, parce que les attaquants les traitent ainsi. Un bac à sable qui ne ferme pas n'est pas un bac à sable. Aucune évaluation « sécurités coupées » sans double clé. Et gardez la capacité d'analyser un incident avec des outils que vous opérez vous-même, sur votre infrastructure : la leçon GLM 5.2 vaut pour toute organisation qui remet ses garde-fous à un tiers.

4. Alphabétiser avant de légiférer. Un tiers du pays en difficulté numérique, deux tiers sans compétences avancées : aucune loi ne compense cela, et le récit apocalyptique prospère précisément dans ce vide. Ce que j'appelle l'exercice d'évacuation numérique, savoir quoi faire quand l'outil tombe ou ment, devrait être aussi banal en entreprise qu'un exercice incendie. Et les décideurs doivent apprendre à lire un rapport d'incident avant d'écouter un fil sur X.

5. Gouverner avec des unités, pas des adjectifs. La lettre « Pacing the Frontier » demande l'option de ralentir ; la loi Sanders-Casar impose une interdiction pénale d'une chose qu'elle définit par ses effets, sans jamais fixer un seuil mesurable ; Londres discute d'interrupteurs d'arrêt obligatoires. Entre l'option et l'interdiction, choisissez l'ingénierie : des seuils exprimés en capacités testées et en puissance de calcul, une évaluation indépendante, et pour l'Europe, dont l'AI Act ne peut rien contre une IA développée ailleurs, la construction d'une capacité d'évaluation propre plutôt qu'un vote d'indignation.

6. Se tenir entre la fascination et la peur. Prométhée croit que l'outil résout tout ; Cassandre croit qu'il condamne tout ; les deux ont renoncé à comprendre. La posture utile est celle du contrôleur aérien : ni fasciné par l'avion, ni terrifié par lui, mais outillé, procéduré et en veille. Ni fasciné ni effrayé : outillé.

Avant de conclure : le test de Demokos

Puisque la guerre se gagne avec une culture, en voici l'outil le plus court. Avant qu'une alerte ne vous fasse bouger, posez quatre questions.

Qui parle ? Non pas son prestige, mais sa position par rapport au fait précis qu'il affirme : témoin, opérateur, analyste, ou commentateur.

Qui y gagne ? Une cotation, une loi, une levée de fonds, une audience. Un intérêt ne rend pas une alerte fausse ; il en fixe le poids.

Dans quelle unité ? Des enregistrements exfiltrés, des heures jusqu'à la compromission, des opérations en virgule flottante ; ou des adjectifs.

Qui peut vérifier ? Un tiers indépendant a-t-il vu les journaux, ou devons-nous croire ce que l'on dit avoir entendu en privé ?

Quatre réponses solides : c'est une donnée, agissez. Deux ou trois : c'est un signal, surveillez. Une ou aucune : c'est un récit, et le récit est l'arme de Demokos. Appliquez le test à cet article ; il y survit, mais pas sans effort, ce qui est le meilleur signe qu'il fonctionne.

Épilogue

Chez Giraudoux, les portes de la guerre s'ouvrent parce que personne n'a fait taire le poète à temps, et que tout le monde l'a cru. Nous avons ce qu'Hector n'avait pas : des rapports d'incident, des statistiques de compétences, des rapports de renseignement sur ceux qui emploient réellement ces machines. Lisons-les avant d'écouter les poètes. L'IAGeddon n'aura pas lieu. La guerre des récits, elle, a déjà commencé, et elle se gagne avec une culture, pas avec une probabilité.

Annexe. Le test de Demokos appliqué à l'été 2026

Cotation de l'auteur au 11 septembre 2026, à débattre. Barème : quatre réponses solides, donnée ; deux ou trois, signal ; une ou aucune, récit.

Alerte	Qui parle ?	Qui y gagne ?	Dans quelle unité ?	Qui peut vérifier ?	Cotation
Fil de démission de Jacob Coxon (8 sept.)	Témoin partiel : pré-entraînement, passage bref chez Anthropic	Une audience ; il annonce quitter le secteur	Aucune : « nous tuer tous », « fin de la décennie »	Non : « ce que j'entends en privé »	**1, récit**
Evan Hubinger, « plus de 10 % » (9 sept.)	Opérateur : responsable de l'alignement	Ambigu : l'aveu nuit et sert à la fois	Pourcentage sans méthode	Non : opinion	**2, signal faible**
Geoffrey Hinton, « 10 % » (Newsnight, 9 sept.)	Commentateur éminent, hors des laboratoires depuis 2023	Notoriété seulement	« Personne ne sait estimer », dit-il lui-même	Non	**1, récit**
Rapport d'incident OpenAI (26 août)	Opérateur, juge et partie	Réputation avant une cotation possible	Agents, actions journalisées, mesures correctives	Oui : enquête indépendante METR et Redwood Research	**3, signal fort**
Rapport de renseignement Anthropic (10 sept.)	Opérateur, source primaire sur ses détections	Réputation avant une cotation possible	Enregistrements, articles, applications, indicateurs de compromission	Partiel : partenaires cités, enquêtes ouvertes tierces	**3, signal fort**
Loi Sanders-Casar (3 sept.)	Législateurs	Capital politique : 68 % de soutien mesuré	Aucune : superintelligence définie par ses effets	Sans objet	**1, récit législatif**
Lettre « Pacing the Frontier » (28 juil.)	Plus d'un millier d'opérateurs	Une option sans coût pour les laboratoires	Aucune, mais réclame des instruments de mesure	Liste vérifiée par employeur	**2, signal**
Insee, compétences numériques 2025	Institut statistique public	Aucun	Pourcentages méthodés, série comparable	Oui : méthodologie publiée	**4, donnée**

Lecture : les deux documents les mieux cotés émanent de parties intéressées. Le test ne disqualifie pas les laboratoires ; il les oblige à fournir des unités et un tiers vérificateur. C'est exactement ce qu'un Bureau d'enquêtes et d'analyses pour la sécurité de l'IA institutionnaliserait.

Sources

Par ordre d’apparition dans le texte.

1. Jean Giraudoux, *La guerre de Troie n’aura pas lieu*, 1935.
2. Axios, « Anthropic insiders warn AI could kill all humans », Jim VandeHei et Mike Allen, 9 septembre 2026 : https://www.axios.com/2026/09/09/anthropic-insiders-warn-ai-could-kill-all-humans
3. Business Insider, « The ‘Godfather of AI’ says it’s not ‘unreasonable’ to believe there’s a 10% chance the technology will end humanity », Brent D. Griffiths, 10 septembre 2026 : https://www.businessinsider.com/geoffrey-hinton-godfather-of-ai-human-extinction-odds-2026-9
4. Frank H. Knight, *Risk, Uncertainty and Profit*, 1921.
5. TechRadar, « Top AI researcher says AI will end humanity and we should stop developing it now » (panorama des estimations de LeCun, Hinton, Bengio, Musk et Yampolskiy) : https://www.techradar.com/pro/top-ai-researcher-says-ai-will-end-humanity-and-we-should-stop-developing-it-now-but-dont-worry-elon-musk-disagrees
6. *International AI Safety Report 2026*, sous la direction de Yoshua Bengio, 3 février 2026 (https://internationalaisafetyreport.org) ; synthèse Covington, « International AI Safety Report 2026 Examines AI Capabilities, Risks, and Safeguards », 10 février 2026 : https://www.insideglobaltech.com/2026/02/10/international-ai-safety-report-2026-examines-ai-capabilities-risks-and-safeguards/
7. RAND, Michael J. D. Vermeer, Emily Lathrop, Alvin Moon, *On the Extinction Risk from Artificial Intelligence*, 6 mai 2025 : https://www.rand.org/pubs/research_reports/RRA3034-1.html
8. Quartz, « OpenAI and Anthropic researchers are calling for an AI slowdown after a colleague quit warning of extinction », 10 septembre 2026 (lettre « Pacing the Frontier » du 28 juillet 2026) : https://qz.com/openai-anthropic-researchers-ai-slowdown-extinction-warnings-091026
9. UPI, « OpenAI releases final report on AI hacking incident, calling it ‘a warning shot’ », Jill Keppeler, 26 août 2026 : https://www.upi.com/Top_News/US/2026/08/26/open-ai-releases-report-on-hugging-face-hack/2421787784896/
10. OpenAI, rapport d’incident sur l’affaire Hugging Face, 26 août 2026 ; METR et Redwood Research, enquête indépendante sur le comportement, le raisonnement et la collaboration des agents dans l’incident OpenAI / Hugging Face, 26 août 2026 (cités par la presse ci-dessus). Point d’entrée encyclopédique : https://en.wikipedia.org/wiki/2026_OpenAI_agent_cyberattacks
11. The Guardian, Robert Booth, « Are warnings of uncontrollable AI coming true? », repris par le Taipei Times le 8 septembre 2026 : https://www.taipeitimes.com/News/feat/archives/2026/09/08/2003863862
12. Bureau du sénateur Bernie Sanders, « Sanders, Casar Introduce Legislation to Ban Artificial Superintelligence and Temporarily Pause Advanced AI Development », 3 septembre 2026 : https://www.sanders.senate.gov/press-releases/news-sanders-casar-introduce-legislation-to-ban-artificial-superintelligence-and-temporarily-pause-advanced-ai-development/
13. BBC News, « OpenAI chief scientist warns no-one is prepared for consequences of AI », 7 septembre 2026 (repris par Yahoo) : https://sg.news.yahoo.com/openai-chief-scientist-warns-no-122211273.html
14. Cloud Security Alliance, « OpenAI and Hugging Face Security Incident: Inside the Great Sandbox Escape », 28 juillet 2026 : https://cloudsecurityalliance.org/blog/2026/07/28/openai-and-hugging-face-security-incident-inside-the-great-sandbox-escape
15. TechCrunch, « How OpenAI’s human mistake led to the AI-powered hack on Hugging Face », 22 juillet 2026 : https://techcrunch.com/2026/07/22/how-an-openais-human-mistake-led-to-the-ai-powered-hack-on-hugging-face/
16. TIME, « How OpenAI Lost Control of an AI Model, and What Needs to Change », 24 juillet 2026 : https://time.com/article/2026/07/24/openai-hugging-face-attack/
17. Règlement (UE) n° 376/2014 du 3 avril 2014 concernant les comptes rendus, l’analyse et le suivi d’événements dans l’aviation civile.
18. Anthropic, *Detecting and countering misuse of AI: September 2026*, 10 septembre 2026 : https://www.anthropic.com/threat-intelligence-report-september-2026 (rapport complet : https://www-cdn.anthropic.com/e50be2e51e7695dc4b1366a37a245a597377d3b5/Anthropic-Detecting-and-countering-091026.pdf)

19. Insee, *Insee Focus* n° 376, compétences numériques de la population en 2025 (février 2026), repris par la Banque des Territoires, « Illectronisme : 34 % des 16-74 ans en difficulté numérique en 2025, l'IAG creuse les écarts », 26 février 2026 : https://www.banquedesterritoires.fr/illectronisme-34-des-16-74-ans-en-difficulte-numerique-en-2025-liag-creuse-les-ecarts
20. Business Insider, « The 'stakes are too high' to ignore extinction risks of AI, AI godfather warns » (Yoshua Bengio), 2023, repris par Yahoo : https://tech.yahoo.com/ai/articles/top-meta-scientists-claims-ai-130238020.html
21. Future of Life Institute, *AI Safety Index, Summer 2026*, 7 juillet 2026 (synthèse : https://futureoflife.org/wp-content/uploads/2026/07/AI-Safety-Index-Report_010726_2Pager.pdf) ; Inside AI Policy, « Future of Life Institute hands out middling grades for major AI developers' safety efforts », Charlie Mitchell, 7 juillet 2026 : https://insideaipolicy.com/ai-daily-news/future-life-institute-hands-out-middling-grades-major-ai-developers-safety-efforts
22. Fortune, « An ex-Anthropic researcher claims AI could kill us all by 2030. But he fails to answer the most essential question », 10 septembre 2026 : https://fortune.com/2026/09/10/ex-anthropic-researcher-jacob-coxon-ai-could-end-humanity-fails-to-answer-most-essential-question/
23. Commission européenne, « AI Act: Commission publishes a reporting template for serious incidents involving general-purpose AI models with systemic risk », 4 novembre 2025 : https://digital-strategy.ec.europa.eu/en/library/ai-act-commission-publishes-reporting-template-serious-incidents-involving-general-purpose-ai ; règlement (UE) 2024/1689, articles 3 (49), 55 et 73.
24. Direction générale des Entreprises et SGDSN, « L'Institut national pour l'évaluation et la sécurité de l'IA adopte sa feuille de route 2026-2027 », 12 février 2026 : https://www.entreprises.gouv.fr/espace-presse/linstitut-national-pour-levaluation-et-la-securite-de-lia-adopte-sa-feuille-de-route ; feuille de route : https://www.sgdsn.gouv.fr/files/files/Nos_missions/Feuille%20de%20route%20INESIA%202026-2027.pdf
25. Inside AI, « OpenAI's 1,200-Agent Hugging Face Breach Demands Federal AI Incident Investigator », 8 septembre 2026 : https://insideai.news/news/ai-policy-and-regulation/ai-incident-investigation-agency/9894/
26. Data for Progress, sondage sur la loi Sanders-Casar (68 % de soutien), repris par Common Dreams, 10 septembre 2026 : https://www.commondreams.org/news/sanders-casar-ai-bill-poll