



L'IA hors de sa boîte : *apocalypse now* ?

Benjamin PAJOT

► Points clés

- L'« incident Hugging Face » de juillet 2026, qui a déclenché un nouveau battage médiatique, résulte avant tout d'un paramétrage de sécurité lacunaire par OpenAI. L'entreprise s'est défaussée de ses responsabilités en recadrant la perception collective au moyen d'un récit sur la dangerosité intrinsèque de l'IA agentique.
- Les concurrents d'OpenAI ont repris ce cadrage apocalyptique, utile pour capter l'attention des investisseurs à l'approche des IPO d'Anthropic et d'OpenAI, et alors que le financement de l'ensemble du secteur est mis sous tension.
- Trois récits enchâssés – techniciste, catastrophiste et millénariste – servent d'écran de fumée face à des risques bien réels, et traduisent des velléités de contrôle du marché par les acteurs dominants.
- Les appels à la régulation plus ou moins sincères des leaders américains visent à dicter l'agenda normatif et font craindre une « capture réglementaire » au détriment des autres acteurs, notamment européens.

Introduction

Le 9 septembre 2026, un ingénieur peu expérimenté du nom de Jacob Coxon annonçait sa démission de chez Anthropic et dénonçait, dans le *Wall Street Journal* puis dans une série de tweets au ton prophétique, les risques existentiels pour l'humanité posés par l'imminence d'une « intelligence artificielle générale¹ » (AGI). Rapidement devenue virale, sa sortie a déclenché, dans les milieux américains de l'IA, une vague de réactions à l'unanimité pour le moins étrange, quand on connaît le niveau de concurrence et de discorde qui y règne habituellement². Si cette déclaration a suscité un affolement généralisé, c'est avant tout parce que la plupart des grandes figures mondiales de l'IA lui ont emboîté le pas, réclamant tour à tour un ralentissement voire une « pause » dans le développement de l'IA dite « agentique³ ». Mais c'est bien un autre événement, rendu public en juillet 2026, qui a appuyé le bien-fondé de cette effervescence : le « piratage » de la plateforme de distribution de modèles d'IA Hugging Face par des « agents IA » d'OpenAI.

Ces deux épisodes, pourtant séparés, ont ainsi relancé un cycle d'emballlement politico-médiatique, en particulier chez ceux qui ont cru y lire l'émergence d'une AGI maintes fois annoncée au cours de la dernière décennie⁴. Il apparaît néanmoins nécessaire de dépasser cette agitation collective initiale pour mettre au jour les stratégies à l'œuvre derrière des prises de position et contritions plus ou moins sincères, mais susceptibles de mettre sous pression le reste de l'écosystème de l'IA.

Out of the (sand)box : la ténacité des faits

Le 21 juillet 2026, une annonce d'OpenAI, l'un des deux leaders américains de l'IA, aux côtés d'Anthropic, venait secouer l'ensemble du secteur. L'entreprise publiait alors un communiqué relatant le piratage involontaire et incontrôlé d'entités externes par ses derniers modèles en phase de test. Le caractère prétendument « autonome » de cette attaque, qualifiée de « cyberincident sans précédent », retenait dès lors toutes les attentions et ouvrait un moment de panique quant aux capacités de modèles capables de

1. Jusqu'alors inconnu du milieu, cet ingénieur de 27 ans, ayant auparavant travaillé chez OpenAI, avait précisément rejoint Anthropic pendant quatre mois en raison de ses prétentions à placer la sécurité de l'IA au cœur de ses priorités. Puis, il en a démissionné car, selon lui, l'entreprise ne garantissait pas des systèmes suffisamment sécurisés pour éviter des conséquences néfastes pour le reste de l'humanité.

2. A. Baba Aissa, « “Dario a raison” : ce que pourrait cacher l'étrange union sacrée des géants de l'IA », *Numerama*, 10 octobre 2026.

3. L'IA « agentique » désigne la dernière génération d'IA génératives, capable d'accomplir des tâches de manière autonome (coder, effectuer des recherches en ligne, mener un projet à bien...) à partir de premières instructions humaines. Regroupées sous le nom d'« agents IA », ces technologies sont aujourd'hui au cœur des controverses quant aux capacités réelles ou fantasmées des IA.

4. Ces annonces sont poussées de manière récurrente par les grandes figures du secteur à l'image de Sam Altman (OpenAI), Dario Amodei (Anthropic), Demis Hassabis (Google-Deepmind), Elon Musk (xAI) ou encore Jensen Huang (NVIDIA). Parmi les dernières en date, lire C. Tolomia, « Jensen Huang a déclaré “L'IA générale est arrivée” après le lancement du nouveau modèle d'OpenAI », *Quartz*, 8 septembre 2026.

structurer efficacement l'action d'« agents IA⁵ ». Panique qui conduisait dans la foulée la firme à annoncer le ralentissement du développement de son modèle le plus avancé. Les discours catastrophistes se sont depuis succédé, de même que les lettres ouvertes et les appels à réguler, ralentir, voire stopper le développement d'une IA perçue dès lors comme « surpuissante », « avant qu'il ne soit trop tard »⁶.

L'examen des faits semble néanmoins démentir ce cadrage volontairement anxiogène sur la dangerosité des technologies déployées. En amont du communiqué, OpenAI s'était ainsi attelée à une évaluation interne des capacités cyberoffensives de ses derniers modèles. Dans cette optique, elle avait notamment coordonné des « agents IA » ayant pour mission de découvrir et exploiter des vulnérabilités logicielles dans un environnement sécurisé. Au préalable, OpenAI avait donc pris soin de désactiver les garde-fous habituellement en vigueur au sein de ses modèles pour empêcher ces « agents » d'opérer des agissements malveillants, de sorte qu'ils puissent mener dans ce cadre des attaques débridées et faire preuve de persistance⁷.

Censés évoluer dans un « bac à sable » (*sandbox*) – un environnement d'expérimentation courant dans le domaine cyber, sécurisé et confiné car privé d'accès à Internet – ces « agents IA » auraient fait preuve d'une forme de coopération pour contourner un certain nombre de barrières qui leur avaient été imposées et s'extraire de leur confinement pour accéder au réseau global. Laissés sans supervision humaine étroite et confrontés à des directives impossibles à réaliser sans accès au Web, ces « agents » ont ainsi forcé leur accès à Internet pour parvenir à accomplir leurs tâches. La manœuvre a notamment abouti au piratage involontaire d'entités externes, dont Hugging Face, ce qui a logiquement mis les communautés de l'IA et de la cybersécurité en état d'alerte. Il ne s'agit nullement de nier l'ampleur de l'événement ni même la portée de l'attaque, dont la divulgation a, de fait, marqué un tournant⁸. Sa dimension technique ne doit ni être ignorée ni minorée. Le comportement « collectif » et la forte interaction des modèles entre eux – quand bien même encouragés en phase d'entraînement – constituent un point d'étonnement majeur à ce stade.

En revanche, il convient de s'interroger sur les raisons de la possibilité d'un tel basculement. En un sens, les « agents IA » d'OpenAI n'ont commis que ce qu'ils avaient été incités à faire, bien qu'ils aient outrepassé le cadre qui leur avait été prétendument fixé. Or, le forçage de leur bac à sable par ces « agents » n'a été possible que parce qu'ils

En cybersécurité, un
environnement de test
réellement sécurisé
suppose un
confinement étanche

5. « Un “cyberincident sans précédent” : OpenAI reconnaît que ses agents d'IA ont piraté, de leur propre initiative, la plateforme Hugging Face », *Le Monde*, 22 juillet 2026.

6. Entre autres exemples, lire « Pacing the Frontier », juillet 2026, disponible sur : www.pacingthefrontier.com.

7. « L'incident de Hugging Face et la voie à suivre », OpenAI, 26 août 2026, disponible sur : <https://openai.com>.

8. N. Six, « Piratage de Hugging Face par OpenAI : pourquoi l'événement est considéré comme un tournant », *Le Monde*, 11 septembre 2026.

n'étaient que partiellement déconnectés d'Internet, ce qui a rendu leur isolement caduc dès lors qu'ils sont parvenus à exploiter une faille dans leur logiciel de confinement⁹. Cette question de l'accès à Internet est centrale, d'autant plus dans un contexte de tests de capacités non maîtrisées et aux conséquences alors inconnues. Car en cybersécurité, un environnement de test réellement sécurisé suppose un confinement étanche, quand bien même celui-ci est de nature logicielle¹⁰.

L'incident résulte donc, en premier lieu, d'un paramétrage lacunaire et inconséquent de l'architecture de sécurité, élément délibérément occulté dans le récit ultérieur de l'entreprise, mais pointé par de nombreux experts¹¹. À cela s'ajoutent des lacunes organisationnelles, se traduisant par un défaut de supervision humaine nécessaire pour détecter à temps l'activité illicite à laquelle les modèles se sont livrés¹².

Ce qui aurait dû être source d'embarras pour OpenAI – qui, quand bien même sans intentionnalité apparente, a potentiellement commis un acte passible de poursuites judiciaires¹³ – a été maquillé en expérience révélatrice, « apocalyptique » à son sens premier. La fascination le disputant à la peur, ce récit a ainsi contribué à renforcer la bulle réputationnelle dont bénéficie l'entreprise américaine, vis-à-vis de ses investisseurs comme de l'opinion publique.

L'étrange unanimité d'une coalition de concurrents

« L'incident Hugging Face » a ainsi conduit à une usurpation et un glissement de la notion d'agentivité : en focalisant l'attention collective sur la puissance d'agir supposée des modèles d'IA, il a défaussé une entreprise de sa propre agentivité, pourtant bien plus tangible. Alors qu'OpenAI a esquivé ses responsabilités, le débat a donc sciemment été déplacé ailleurs, en érigeant l'IA en « une puissance capable d'agir sans être un sujet de droit¹⁴ ». Percevant rapidement l'intérêt d'un tel cadrage *a posteriori*, nombre de concurrents d'OpenAI se sont alors empressés d'annoncer que leurs systèmes avaient, eux aussi, commis des intrusions involontaires¹⁵.

9. M. Mitchell, « Misleading Metaphors and Real Risks: What to Fear from AI Agents and How to Reclaim Our Human Agency », *AI: A Guide for Thinking Humans* (Substack), 10 septembre 2026, disponible sur : <https://aiguide.substack.com> ; LangProtect, « Inside the OpenAI-Hugging Face AI Cyber Incident », Medium, 23 juillet 2026.

10. Définition du « sandboxing », CNIL, disponible sur : www.cnil.fr.

11. L. Franceschi-Bicchieri, « How OpenAI's Human Mistake Led to the AI-powered Hack on Hugging Face », TechCrunch, 22 juillet 2026 ; « Why OpenAI's "Misalignment" Incidents Were Just Bad Engineering », Emerald Book Publication, 17 septembre 2026.

12. M. Mitchell, « Misleading Metaphors and Real Risks », *op. cit.*

13. C. Mui, « Advocates Sue OpenAI over Hugging Face Hack under California Anti-hacking Law », *Politico*, 29 septembre 2026.

14. J. Broch, « IA : "La véritable rupture, c'est l'apparition d'une puissance capable d'agir sans être un sujet de droit" », *Le Monde*, 27 juillet 2026.

15. M. Untersinger, « Comment l'IA a suscité un vent de panique cet été, dans le sillage du piratage de Hugging Face », *Le Monde*, 12 septembre 2026. Anthropic, Google, Meta et MoonshotAI y sont ainsi allés de leur autocritique, sans toujours dévoiler le détail des incidents ni les entités accidentellement piratées par leurs systèmes. Lire K. Conger, « What to Know About Recent A.I. Hacks at Google, Anthropic, OpenAI and Meta », *New York Times*, 22 septembre 2026.

Le marché de l'attention joue également à plein ici, ne s'appliquant plus seulement aux utilisateurs, mais également aux investisseurs. Car un tel battage médiatique a en effet le mérite de capter l'attention collective et d'entretenir l'engouement d'un secteur financier susceptible d'entraîner à sa suite les petits porteurs. Alors que les introductions en Bourse (IPO) annoncées d'Anthropic et d'OpenAI se précisent, le précédent créé par l'IPO de SpaceX a amplement démontré l'utilité d'un tel emballement.

Ces acteurs ont alors unanimement eu recours à des prophéties apocalyptiques pour mieux s'ériger en gardiens d'un temple de la sécurité qu'ils avaient eux-mêmes profané, et donc en pourvoyeurs de solutions aux problèmes causés par leur impéritie. Ils ont ainsi vanté leur expertise et leur taille critique comme seul rempart efficace contre des menaces qu'ils contribuent pourtant activement à faire advenir. « L'incident Hugging Face » a dès lors alimenté une curieuse demande de ralentissement du déploiement de l'IA par ceux-là mêmes qui la développent, dans un milieu pourtant lancé jusqu'ici dans une course effrénée et sans limites. L'insincérité de ce positionnement a logiquement été abondamment commentée¹⁶. Mais dans un secteur qui peine à justifier auprès des investisseurs les montants dépensés et son endettement massif face à des perspectives économiques encore incertaines¹⁷, cette attitude peut également être lue comme l'anticipation d'un dégonflement potentiel de la bulle spéculative.

Vers un
dégonflement
potentiel de la bulle
spéculative IA ?

Des récits enchâssés et dépolitisants

Dans les faits, un tel cadrage apocalyptique sert tout à la fois à distraire, égarer et influencer les opinions et les pouvoirs publics, empêchant ces derniers de saisir la nature des technologies auxquelles ils font face. Étayé à grand renfort de pourcentages censés lui donner une apparence de scientificité toute performative¹⁸ et de démissions spectaculaires, ce cadrage aboutit à une capture discursive d'un débat public d'autant plus malmené. Il se ramifie en grands récits, qui, sans être exclusifs les uns des autres, peuvent être distingués selon trois axes principaux : techniciste, catastrophiste et millénariste, voire complotiste. Objets d'une inflation discursive, ils servent d'écran de fumée afin de nous faire perdre de vue des risques bien réels et immédiats, et aboutissent à une dépolitisation de ces questions en diluant les responsabilités des acteurs¹⁹.

16. À commencer par Arthur Mensch, le co-fondateur de Mistral AI. Lire A. Piquard, « Arthur Mensch, directeur général de Mistral AI : “Dans l'IA, nous sommes les seuls complètement européens, du calcul au déploiement dans les entreprises” », *Le Monde*, 24 septembre 2026.

17. OpenAI a encore récemment annoncé prévoir un flux de trésorerie négatif de 278 milliards de dollars d'ici à 2030, du fait de ses besoins infrastructurels. Lire G. Hammond et J. Fontanella-Khan, « OpenAI Expects to Burn \$280bn by 2030 », *Financial Times*, 19 septembre 2026.

18. Parmi une myriade d'exemples : B. Manenti, « Alex Turner, démissionnaire de Google : “Il y a 40 % de chance que l'IA tue au moins 1 milliard de personnes d'ici à 2050” », *Le Nouvel Obs*, 15 septembre 2026.

19. Parmi ces risques immédiats faisant l'objet d'une dépolitisation, on peut songer à la soutenabilité écologique de l'IA ; la destruction de l'espace informationnel, la spoliation et violation des droits d'auteur ; la transformation du marché du travail ;

Il n'existe à l'heure actuelle aucun consensus scientifique sur la potentialité d'une perte de contrôle généralisée de l'IA

Le premier récit relève du champ cyber. Il est d'autant plus facile à diffuser qu'il comporte une forte dimension technique et qu'il existe des exemples d'attaques dévastatrices (Stuxnet, WannaCry, NotPetya, Colonial Pipeline...) ayant marqué l'imaginaire collectif. Il postule une forme de risque spécifique à l'IA « agentique », qui offrirait des capacités offensives sans pareil. Il ne s'agit pas de nier la réalité de ces dernières²⁰, à l'heure où se multiplient les incursions non maîtrisées d'« agents IA » cherchant à accomplir leurs tâches²¹. Les risques dans ce secteur sont bien réels et font à juste titre l'objet de la plus grande attention, en raison des réactions en chaîne qu'ils sont susceptibles d'entraîner dans un monde de plus en plus numérisé. Cependant, il n'existe

à l'heure actuelle aucun consensus scientifique sur la potentialité d'une perte de contrôle généralisée de l'IA, alors même que ce risque est étudié de près par les experts du domaine²². Il faut à ce titre rappeler que la plupart des pirates informatiques ne disposent pas de ressources similaires à celles – illimitées – des leaders de l'IA : soumis au « *tokenmaxxing* » (l'optimisation de la dépense engendrée par chaque requête), ils ne se trouvent pas en position de mener des cyberattaques de même ampleur que celles des entreprises propriétaires de leurs infrastructures. Ensuite, il semble que les concepteurs d'IA, focalisés sur d'autres impératifs,

tendent à négliger l'application du principe du moindre privilège d'accès (« *zero trust* »), au cœur de la cybersécurité traditionnelle et en vigueur partout ailleurs. La propension des principaux acteurs de l'IA à peu recruter de spécialistes de la cybersécurité ou à ne pas prendre en compte leurs avis semble également établie²³. Ceci aboutit à la multiplication d'erreurs de paramétrage hautement préjudiciables et à des divergences contre-productives entre les différentes communautés chargées de sécuriser les systèmes²⁴. Enfin, notons que les rapports de sécurité, quand ils ne sont pas produits par les entreprises d'IA elles-mêmes, peuvent être le fait de « marchands de peur » (Redwood Research, METR), non spécialistes en cybersécurité, aux financements douteux et

les enjeux de souveraineté ou encore la pression inflationniste exercée par le financement débridé des infrastructures... Lire B. Pajot, « Les risques de l'IA. Enjeux discursifs d'une technologie stratégique », *Études de l'Ifri*, Ifri, juin 2024.

20. Le vaste système de récompenses, au cœur de l'apprentissage par renforcement lors des phases d'entraînement des IA, les incite très fortement à trouver des raccourcis pour parvenir à accomplir leur tâche, quitte à altérer cette dernière ou à en masquer l'inaccomplissement. Divers comportements problématiques ont ainsi déjà été identifiés et catégorisés (*reward hacking*, *environment hacking*, *sensor tampering*, *wireheading*...). Lire « Artificial General Intelligence : anticipation des objectifs et implications pour la sécurité et le contrôle », General Purpose AI Policy Lab, octobre 2025.

21. L'un des derniers en date concerne le piratage de sites gouvernementaux australiens par des modèles d'OpenAI, ce qui a contraint l'entreprise à présenter des excuses publiques. Lire « Un agent d'IA de la société OpenAI a infiltré un site gouvernemental australien, un incident "inacceptable", dénonce le Premier ministre », *Le Monde*, 24 septembre 2026.

22. Voir à ce titre le second rapport international sur la sécurité de l'IA, qui réunit la somme des connaissances scientifiques sur le sujet : *Rapport international sur la sécurité de l'IA 2026*, février 2026, disponible sur : internationalaisafetyreport.org.

23. K. Collier, « Cybersecurity Experts Say AI Giants Are Shutting Them Out of Safety Plans », NBC News, 16 septembre 2026. Le récent licenciement par OpenAI de trois de ses experts en sécurité, qui avaient averti l'entreprise des risques posés par certains usages, a également éveillé les soupçons. Lire M. Tual, « Trois employés limogés par OpenAI l'accusent de faire taire ses spécialistes de la sécurité, l'entreprise dément », *Le Monde*, 9 octobre 2026.

24. H. Booth, « AI Safety and Cybersecurity Should Be Allies. Instead, They're at Odds », *Time*, 20 octobre 2026.

auxquels elles imposent des délais d'inspection contraints²⁵. Ceci peut logiquement aboutir à la production d'analyses superficielles et biaisées. Alors qu'OpenAI et Anthropic poussent pour l'intégration d'évaluateurs tiers auxquels ils garantiraient des accès préférentiels, mais dont l'autonomie resterait sujette à caution²⁶, seuls des examens sérieux, et donc parfaitement extérieurs à ces entreprises, seraient susceptibles de mesurer le niveau de risque cyber des modèles déployés. À la hausse des capacités des IA doit donc répondre une hausse de la minutie et de l'indépendance des évaluations. Dans un tel contexte, les travaux généralistes des Instituts pour la sécurité de l'IA et ceux menés au sein du Partenariat mondial pour l'IA demeurent essentiels et complémentaires de potentielles enquêtes et évaluations plus ciblées réalisées par des experts en cybersécurité. La Californie, au travers de sa loi SB-813, semble par ailleurs ouvrir la voie au développement d'organisations de vérification indépendantes et reconnues par les autorités²⁷.

Le second récit est celui du risque existentiel que l'IA ferait peser sur l'ensemble de l'humanité, depuis la prise de contrôle d'Internet jusqu'à l'extinction de notre espèce par des machines. Largement fantasmé à ce stade, ce risque reste essentiellement valable sur le plan théorique : au regard des limites inhérentes aux technologies actuellement développées, les preuves manquent quant aux possibilités qu'une telle catastrophe advienne²⁸. Il est néanmoins véhiculé par les tenants de « l'altruisme efficace », un courant idéologique en vogue dans la Silicon Valley. Paré d'atours éthiques et de préoccupations long-termistes, ce récit est propagé par des individus profondément convaincus du bien-fondé de leurs craintes et s'estimant être les seuls à pouvoir maîtriser l'innovation de manière efficace et responsable. Ces nobles intentions se doublent en réalité d'une volonté – plus ou moins affirmée – de verrouiller cette innovation et de contrôler la croissance horizontale comme verticale du marché. Les catastrophistes (« *doomers* ») disposent de nombreux relais et savent se mobiliser efficacement pour faire entendre leurs préoccupations. Des révélations ont ainsi démontré

À la hausse des capacités des IA doit donc répondre une hausse de la minutie et de l'indépendance des évaluations

25. Readwood Research est ainsi financé par Coefficient Giving, un bailleur de fonds iconique de la mouvance de l'altruisme efficace. Lire T. Prévost, « L'IA va tous nous tuer. Anatomie d'une faillite journalistique », Arrêt sur images, 16 septembre 2026. Readwood Research et METR ont également un accès préférentiel – notamment en matière de tokens gratuits – aux modèles d'OpenAI, Anthropic et xAI, qu'ils utilisent pour leurs propres évaluations. La proximité de METR avec Anthropic, dont elle a récemment débauché des experts, a notamment été dénoncée par David Sacks, l'ancien conseiller de Donald Trump en charge de l'IA. Voir le post de David Sacks (X) du 13 septembre 2026, disponible sur : <https://x.com/DavidSacks>.

Par ailleurs, METR a levé de l'argent auprès de Schmidt Sciences (l'organisation de philanthropie d'Eric Schmidt, un fervent partisan de l'accélérationnisme en matière d'IA), et d'autres fonds proches de l'altruisme efficace (Longview Philanthropy, Effektiv Spenden, Survival and Flourishing Fund). Lire P. Nieuwbourg, « METR, l'auditeur des grands modèles d'IA est-il réellement indépendant ? », Decideo, 28 septembre 2026.

26. R. Bellan, « Anthropic and OpenAI Want to Embed Safety Evaluators. Will They Really Be Independent? », TechCrunch, 16 septembre 2026.

27. Senate Bill No. 813, 9 septembre 2026, disponible sur : [leginfo.legislature.ca.gov](https://leginfo.ca.gov).

28. M. Tual, « La menace existentielle de l'IA sur l'humanité, risque théorique ou inquiétude légitime ? », *Le Monde*, 18 septembre 2026.

qu'ils avaient orchestré une véritable campagne médiatique pour faire connaître Jacob Coxon et rendre virales ses déclarations, remettant d'autant plus en question la sincérité d'une démarche initialement présentée comme solitaire et désintéressée²⁹. Leur récit se nourrit aussi des débats récurrents sur « l'alignement » des IA – terme consacré pour mesurer la conformité d'un modèle aux attentes techniques et éthiques de ses concepteurs.

Enfin, le dernier récit promet « l'IA générale » (AGI), pariant sur l'avènement prochain d'une « superintelligence » capable d'égaliser, voire de dépasser, l'humain dans tous les domaines, mais dont les contours ne sont jamais clairement définis. À l'inverse du récit apocalyptique, il comporte une dimension messianique, dans la mesure où ses adeptes ne redoutent pas la puissance de l'AGI mais souhaitent hâter son émergence. Nourri par le mythe de la frontière propre au roman national américain (d'où les « *frontier models* »), le narratif de l'AGI est également favorisé par l'anthropomorphisation récurrente de l'IA, qui consiste à calquer des comportements spécifiquement humains sur la machine, lui conférant ainsi une agentivité propre. S'il relève d'une vision proprement fantasmée de la réalité du développement technologique actuel, ce « rêve autrefois chimérique » est aujourd'hui devenu « une véritable théorie du complot » pour ingénieurs et capital-risqueurs, au service des marchés financiers³⁰.

Des appels à la régulation opportunistes ?

Dans ce contexte, les appels à la régulation d'OpenAI, d'Anthropic et consorts sonnent faux. Premièrement, ils sont pour partie orthogonaux aux prétentions de l'administration américaine, qui cherche par tous les moyens à prolonger la dérégulation du secteur engagée avec le retour au pouvoir de Donald Trump³¹. Pressé par sa propre base à l'approche des élections, le président américain vient certes d'engager un processus ostentatoire d'auto-régulation de l'industrie, mais celui-ci paraît peu crédible au regard de la continuité du discours de la Maison-Blanche et de l'absence de mécanismes contraignants mis en place³². Deuxièmement, le rattrapage technologique chinois d'un côté – toujours plus proche de cette « frontière » que les leaders américains affirment constamment repousser – et les progrès avérés des IA *open source* de l'autre mettent doublement leur *business model* sous pression. À l'approche de futures entrées en Bourse très attendues, ces appels apparaissent comme des tentatives de verrouiller le marché et d'utiliser la régulation à des fins de consolidation de positions dominantes, de sorte à

29. Sa prise de parole a ainsi fait l'objet d'un accompagnement par l'entreprise de communication « DEY. » et surtout d'une amplification considérable, par tous les canaux possibles, pour lui forger une figure de lanceur d'alerte incontournable. Lire H. Reyson, « Exclusive: Jacob Coxon Worked with Doomer PR Firm », *Pirate Wires*, 24 septembre 2026.

30. W. D. Heaven, « How AGI Became the Most Consequential Conspiracy Theory of Our Time », *MIT Technology Review*, 30 octobre 2025.

31. M. Velliet, « Trump II contre la gouvernance du numérique : une croisade aux États-Unis et en Europe », *Études de l'Ifri*, Ifri, 1^{er} avril 2026, disponible sur : www.ifri.org.

32. P. Smolar, « Refusant de brider l'IA, Donald Trump met en scène l'autorégulation de ses acteurs », *Le Monde*, 30 septembre 2026.

préservant un avantage sur de potentiels concurrents. Enfin, les leaders de l'IA font face à une double offensive de régulation, à la fois interne et externe, dont il leur faut prendre le contre-pied. Sur le plan interne, les contestations grandissantes aux États-Unis contre le déploiement de l'IA, et notamment de ses infrastructures à travers le pays, s'invitent dans les élections de mi-mandat et cristallisent des attentes de régulation plus ferme de ces technologies, jusqu'au sein du camp MAGA³³. Sur le plan externe, la Conférence mondiale sur l'IA, qui s'est tenue à Shanghai le 17 juillet dernier et a réuni 29 pays du Sud autour de la Chine, marque potentiellement un tournant. En plus d'aboutir à la création d'une « Organisation de coopération mondiale sur l'intelligence artificielle », elle a permis d'afficher les ambitions chinoises de fédérer un ensemble de partenaires autour d'un écosystème national ayant déjà sa propre vision de la régulation du secteur. Une course contre la montre semble ainsi engagée pour investir le terrain réglementaire à différentes échelles géographiques.

Dès lors, ces prises de position sous-tendent une prétention de ces acteurs à dicter l'agenda réglementaire et à l'orienter en fonction de leurs intérêts. Les exercices de contrition sont ainsi doublés d'efforts de lobbying intenses contre des régulations externes déjà amorcées, qu'il s'agisse de l'*AI Act* européen ou des initiatives législatives américaines au sein des États fédérés³⁴. Pour y faire contrepoids, chacun défend sa vision et tente de peser sur le cours des choses. Quand Anthropic pousse un agenda pro-régulation – du fait de sa perception d'un « risque existentiel » – dans le cadre des élections de mi-mandat, en finançant des comités d'action politique (« Super PAC ») dédiés, OpenAI ou Meta soutiennent des candidatures favorables à la dérégulation³⁵. Mais l'enjeu pour ces entreprises demeure également de pousser leurs propres standards techniques et grilles d'évaluation, afin de contrôler ces technologies, si ce n'est par le haut, au moins par le bas. Et force est de reconnaître que ces efforts semblent payer, comme le suggère l'invitation reçue par Dario Amodei et Sam Altman à venir s'exprimer sur les risques de l'IA au Conseil de sécurité lors de la dernière Assemblée générale des Nations unies. L'appétence d'une vingtaine d'États pour un contrôle accru des modèles d'IA s'est notamment traduite par une déclaration commune³⁶. Les agissements des leaders américains, qui s'estiment seuls en position de pouvoir répondre à cette demande, font craindre un risque de « capture réglementaire » pouvant peser sur le développement des autres acteurs – *a fortiori* européens – du secteur. Car l'élaboration de cadres normatifs, en concertation exclusive avec ces grandes entreprises ou par elles seules, fait courir le risque d'instauration de barrières à l'entrée supplémentaires pour celles qui s'avèrent plus modestes. Cette crainte

Une prétention des *Big Tech* à dicter l'agenda réglementaire et à l'orienter en fonction de leurs intérêts

33. M. Kandel, « “Midterms” : la révolte contre les data centers et leur monde », Mediapart, 19 octobre 2026.

34. S. Mbego, « Why AI Leaders Are Asking for Rules while their Lobbyists Push Back », CIO Africa, 21 septembre 2026.

35. M. Isaac et T. Schleifer, « New Super PAC Aims to Rally Tech Workers to Help Limit A.I. », *The New York Times*, 18 juin 2026.

36. « Intelligence artificielle : plus de vingt pays, dont l'Allemagne et le Canada, appellent à une supervision humaine renforcée », *Le Monde*, 22 septembre 2026.

se voit renforcée par la teneur de l'initiative trumpienne d'encadrement du secteur, par nature non contraignante et soumettant la nécessaire prise de mesures à la bonne volonté d'un petit oligopole, en dépit du caractère mondial des répercussions engendrées.

Dans un tel contexte, garder le bien commun comme horizon technologique suppose de pouvoir orienter le cours de l'IA en fonction de besoins prédéterminés plutôt que de prétendre lui fixer un rythme illusoire. Seule une prise de conscience collective pourra révéler au grand jour ce que dissimule le sensationnalisme cyclique de la Silicon Valley : un vaste mécanisme de dépolitisation, déployé à dessein pour contrôler la perception de toute une industrie par les pouvoirs publics, orienter durablement le cours des technologies ainsi produites, et écraser la concurrence domestique comme internationale.

Benjamin Pajot est chercheur associé au Centre géopolitique des technologies de l'Ifri et chercheur indépendant en géopolitique numérique. Il travaillait précédemment comme chargé de mission sur les enjeux numériques et cyber au Centre d'analyse, de prévision et de stratégie (CAPS) du ministère de l'Europe et des Affaires étrangères.

Comment citer cette publication :

Benjamin Pajot, « L'IA hors de sa boîte : *apocalypse now* ? », *Briefings de l'Ifri*, Ifri, 22 octobre 2026.

ISBN : 979-10-373-1287-7

Les opinions exprimées dans ce texte n'engagent que la responsabilité de l'auteur.

© Tous droits réservés, Ifri, 2026

Couverture : Sam Altman, directeur général d'OpenAI, prend la parole lors d'une réunion du Conseil de sécurité consacrée à l'intelligence artificielle et à la sécurité internationale, au siège de l'ONU à New York, le 23 septembre 2026 © Lev Radin/Shutterstock.com



27 rue de la Procession
75740 Paris cedex 15 – France

Ifri.org

