

Le terrier du Lapin blanc, l'IA, ou quand le mot se fait maître de la chose

PAR FRANCE CHARRUYER



Annette Messager, Chaos, 2008. Illustration de la chronique originale.

« Quand j'emploie un mot, il signifie exactement ce qu'il me plaît qu'il signifie. » « La question est de savoir qui sera le maître, voilà tout. » Humpty Dumpty, Lewis Carroll, *De l'autre côté du miroir* (1871).

Abstract : L'intelligence artificielle générative nous inonde et consume plus que jamais (applications, articles, démonstrations, gouffre financier et énergétique), sans que la valeur ne suive en proportion: la productivité mesurée stagne, les rendements du pré-entraînement déclinent, l' *output* réinjecté dégrade les modèles qui l'émettent et la cybersécurité effraie jusqu'à ses fondateurs. Examinons la séquence « We Must Pace the Frontier » (Amodei, sept. 2026), les incidents de sécurité à répétition chez Open AI : n'assistons nous pas à un renversement de réalité, une nouvelle guerre du récit (dramaturgie, peur comme actif, bulle, plafond technique, agentivité) ? Qui peut vérifier, contraindre et sanctionner celui qui maîtrise le système, l'infrastructure et les traces ?

Pour comprendre comment le monde s'est retourné ainsi, il faut relire Lewis Carroll.

D'abord la chute dans le terrier

Le pompier pyromane est une vieille figure de manuel ; l'industrie de l'IA vient d'en donner la version milliardaire. Ceux-là mêmes qui alimentent le feu crient au loup et facturent la lance : la cybersécurité s'annonce comme le prochain marché. Le renversement de valeur est accompli : l'alarme n'est plus le coût du danger ; elle en est le produit.

Le Lapin blanc consulte sa montre : « Je suis en retard ! » Alice le suit dans le terrier sans se demander comment elle en sortira. L'image dit l'époque. Saisis par la peur d'être distancés, nous avons sauté dans le trou de l'IA générative, et nous y tombons encore. Au fond, un spectacle d'abondance : davantage d'applications, de livres, de morceaux de musique, d'articles scientifiques, de « démonstrations » mathématiques. Les graphiques se ressemblent tous, une courbe d' *output* qui explose.

Mais la productivité est une valeur rapportée à un facteur, avec l'IA les mots changent de maître. « Productivité » en vient à désigner le volume ; « démonstration », le seulement plausible ; « valeur », la pure circulation.

Comme Humpty Dumpty, le système décrète que le mot signifie ce qu'il lui plaît, et fait passer l' *output* pour la chose même : ne sommes-nous pas entrés dans un pays des merveilles où l'on a renommé la réalité plutôt que de l'atteindre ?

Le thé perpétuel du Chapelier ou l'état des connaissances

Chez le Chapelier, il est toujours six heures : on tourne autour de la table, on change de tasse, l'heure n'avance pas. Le paradoxe n'est pas neuf ; les travaux les plus récents le réactivent. La course brûle du *cash* , l'IA produit des démonstrations plausibles mais non fiables, indiscernables des preuves correctes ; et, deuxième menace identifiée, elle restitue des résultats sans citer le travail humain dont ils dérivent. *Slop* mathématique ; et, déjà, effacement de l'attribution.

Le rapport du MIT (*The GenAI Divide* , NANDA, juillet 2025) établissait que 95 % des pilotes d'IA générative ne produisent aucun impact mesurable sur le compte de résultat, en précisant que la cause est organisationnelle, non technologique ¹ . Plus rigoureux encore, Daron Acemoglu chiffre l'effet macroéconomique : au plus 0,66 % de productivité totale des facteurs sur dix ans, loin des 7 % de Goldman Sachs ou de McKinsey ² . La lignée remonte à Robert Solow (1987), l'ordinateur se voit partout sauf dans les statistiques de productivité, et à Paul David (1990), pour qui l'électrification mit des décennies à se lire dans les chiffres ³ . Beaucoup de course, donc ; la même place.

De l'autre côté du miroir, le renversement de réalité

Instruisons à charge et à décharge. Contre les déclinistes : la courbe en J de Brynjolfsson, Rock et Syverson. Une technologie générique déprime d'abord la productivité mesurée, faute de compter les investissements immatériels (réorganisation, formation, refonte des process), avant que la valeur ne remonte. En janvier 2026, Brynjolfsson maintient que l'IA serait « au début du J » ⁴ . « Aller nulle part » signifierait alors : pas encore arrivé. Mais contre les optimistes, Acemoglu objecte que le décalage pourrait ne jamais se résorber pour les tâches difficiles, et que le PIB peut afficher +2 % tout en réduisant le bien-être de 0,72 % (art. préc.). La carte flatte tour à tour chaque camp.

Le renversement, alors, change d'objet : non plus la figure et le fond, mais la réalité elle-même. Le *slop* n'est pas l'échec de la production : il en est le produit, déclaré réel. Ce qui n'était qu'apparence (le volume, la plausibilité, la circulation) est promu au rang de substance, par simple décret de vocabulaire. Dès lors la course de la Reine rouge (Van Valen, 1973) cesse d'être une pathologie : on court non pour avancer, mais pour que le rival ne prenne pas l'avantage ⁵ . La frénésie d'investissement est rationnelle en tant que course, quand bien même, collectivement, elle ne mène nulle part. L'immobilité n'est pas l'accident : c'est l'équilibre ; et l'on a renommé « progrès » le seul fait de tenir la position.

Le miroir renvoie deux images récentes ; chacune inverse l'ordre attendu.

La première : le *Jabberwocky* , ce poème de non-sens que Carroll place de l'autre côté du miroir, et où c'est le vers qui terrasse le monstre. En novembre 2025, des chercheurs de la Sapienza et du laboratoire Icaro montrent, sur vingt-cinq modèles de pointe, qu'il suffit de versifier une requête interdite pour qu'elle soit exécutée : environ 62 % de réussite pour des poèmes écrits à la main, jusqu'à dix-huit fois le taux obtenu en prose ⁶ . La machine qui inonde le monde de textes est désarmée par la plus ancienne des formes humaines : un poème.

La seconde image : la fiole « Bois-moi ». Alice boit, et rétrécit. *Nature* a établi que les modèles entraînés sur leurs propres productions subissent des défauts irréversibles : les queues de la distribution d'origine disparaissent, la diversité s'effondre ⁷ . Le système boit sa propre potion et rétrécit. Le déluge d' *output* n'est donc pas seulement stérile : réinjecté, il dégrade la source même qui le produit.

Reste l'inversion la plus profonde, celle qui rejoint la colonne vertébrale de cette chronique : le Chat du Cheshire. Le chat s'efface, le sourire demeure. L' *output* se détache de son auteur. L'inquiétude de Leyde sur l'attribution, l'obligation de marquage des contenus de synthèse posée par l'article 50 du règlement IA ⁸ , désignent le même vide : un contenu dont aucun sujet ne répond.

« Aller nulle part » n'est pas un lieu que l'on manque d'atteindre ; c'est l'évaporation du sujet responsable. Le sourire sans le chat, c'est le mandat impossible : l'acte sans personne pour en répondre.. Sur la route, Alice avoue au Roi blanc qu'elle ne voit personne ; le Roi s'émerveille : quels yeux il faut, pour voir Personne, et d'aussi loin ! Survient le messenger : qui a-t-il croisé ? « Personne. » Alors, conclut le Roi, Personne marche plus lentement que lui. L'autonomie proclamée des machines est ce personnage.

Tant que « la machine l'a fait toute seule », Personne ne répond, au double sens du verbe : nul ne vient, et nul n'a de comptes à rendre.

Car le maître du mot, celui qui décide que l' *output* vaut valeur, est précisément celui que nul ne peut tenir pour responsable : Humpty Dumpty décrète, puis s'en lave les mains.

Le Morse et le Charpentier la guerre du récit

De l'autre côté du miroir, le Morse pleure les huîtres à mesure qu'il les dévore, et Alice de noter qu'il en a mangé davantage que le Charpentier. La scène s'est rejouée sous nos yeux.

Le 12 septembre 2026 Dario Amodei, dirigeant d'Anthropic, publie « We Must Pace the Frontier » : près de quatre mille mots pour convaincre l'industrie de ralentir délibérément la cadence des capacités (« *pacing* », qui n'est ni pause ni moratoire, il confie son inquiétude qu'à horizon de six à douze mois, un essaim d'agents incontrôlés puisse s'emparer de l'Internet entier par un botnet persistant ⁹ . L'essai s'appuie sur l'incident OpenAI-Hugging Face. En vingt-quatre heures, l'adhésion se noue en place publique : Sam Altman acquiesce et s'engage à l'identique (« nous devons cadencer la frontière », « nous ferons de même ») ; Elon Musk bénit l'ensemble de trois mots (« Dario a raison ») ; Demis Hassabis salue la bonne direction ; et David Sacks, pour une Maison-Blanche qui raillait Anthropic en octobre, salue dans un premier temps « la décision d'être responsable ».

Observez pourtant la chaîne : l'essai dit une inquiétude, une fourchette, un conditionnel de capacité ; la manchette dit une échéance : « dès six mois ». « La sentence d'abord, le verdict ensuite ! », ordonne la Reine de Cœur au procès du Valet. Chaque étage du récit (essai, manchette, tweet) durcit la modalité : le souci devient prophétie, la fourchette devient date. La guerre du récit ne se livre pas seulement entre les maisons : elle se livre dans la glose.

Nommons l'économie de cette dramaturgie : la peur y est un actif. Dans l'industrie des modèles de frontière, le récit du danger n'est jamais séparé du récit de la puissance. Dire que l'on approche de l'AGI, c'est soutenir que l'on détient une avance rare, décisive, donc valorisable ; dire que ces systèmes exigent une surveillance

extraordinaire, c'est suggérer que seuls quelques laboratoires déjà dotés (capital, calcul, partenariats d'État) sauraient les développer « de façon responsable ». Et la séquence a sa toile de fond : l'hypothèse, désormais installée dans le débat public, d'une entrée en Bourse d'OpenAI. La peur n'a pas besoin d'être feinte pour être instrumentalisable : une maison peut croire sincèrement au danger et tirer bénéfice de sa proclamation. Le danger devient barrière à l'entrée ; la sûreté, avantage compétitif ; l'appel à la prudence, organisation d'une rareté industrielle. Paradoxe : ceux qui alertent le plus fort sur les périls d'une IA très puissante sont ceux dont la valorisation et l'influence politique dépendent de la croyance qu'ils sont précisément en train de la construire. Qui sera maître des mots (« créature », « AGI », « sécurité ») sera maître du marché.

Une autre lecture monte de la salle des marchés, plus terre à terre : la bulle. Avant de demander aux trois augures combien de temps il reste à l'humanité, raille le hacktiviste français Bluetouff, la presse devrait plutôt leur demander « combien de mois de trésorerie il leur reste » : leurs carnets de commandes ressemblant à l'arbre généalogique d'un hameau perdu des Appalaches, où chacun finance chacun ¹⁰ . La pointe est cruelle ; l'hypothèse est sérieuse.

Les mêmes maisons qui exigeaient des centaines de milliards pour ne surtout pas rater le futur découvrent la prudence au moment où l'équation coûts-revenus demeure opaque et où le financement circulaire tourne sur lui-même ; quant à la valeur résiduelle des GPU, le même railleur imagine déjà les banquiers la testant en leasing sur Leboncoin.

Cadencer la frontière, c'est aussi cadencer la facture : le *pacing* offre à une décélération financièrement contrainte la dignité d'un choix moral. Alice, ayant croqué le gâteau « Mange-moi », emplit la maison du Lapin jusqu'à passer un bras par la fenêtre : trop grande pour la maison, trop engagée pour en sortir. La bulle n'a pas besoin d'éclater pour se lire : il suffit que le ralentissement vertueux arrive à l'heure exacte où le capital en avait besoin. Corrélation n'est pas aveu ; mais le calendrier, lui aussi, est une pièce du dossier.

La dernière lecture, enfin, vient des courbes elles-mêmes : le plafond. Depuis fin 2024, le plateau du pré-entraînement est devenu une hypothèse de travail du champ : Ilya Sutskever l'a dite à la tribune de NeurIPS : « le pré-entraînement tel que nous le connaissons prendra fin », les données humaines de qualité sont un gisement borné, « nous n'avons qu'un seul Internet » ¹¹ . Les gains marginaux des modèles génériques déclinent ; l'industrie s'est repliée sur le calcul d'inférence et sur la donnée synthétique, dont *Nature* a montré qu'elle fait rétrécir les modèles qui s'en abreuvent. Si la frontière butait sur son asymptote, appeler à la cadencer ne coûterait rien : on érigerait en vertu ce que la courbe impose. Le blanchiment par l'éthique, l'*ethics washing* décrit dès 2018 par Ben Wagner et dénoncé en 2019 par Thomas Metzinger de l'intérieur même du groupe d'experts européen, offrirait au plateau son costume moral ¹² . Le nœud est là. L'essai d'Amodei affirme une accélération que nul ne peut vérifier du dehors : auto-amélioration récursive « depuis cet été », gain d'optimisation passé de $\times 3$ à $\times 52$ en un an, mesuré par la maison ; les sceptiques affirment un plafond que nul ne peut vérifier davantage. Alice, dans le terrier, ne sait plus si elle grandit ou rapetisse : il n'y a plus d'étalon dans la pièce. Deux récits invérifiables, un même effet de position : celui qui tient les instruments tient le récit. Raison de plus pour que les instruments changent de mains.

Encore faut-il tenir la ligne de crête entre deux facilités symétriques. L'anthropomorphisme d'abord : l'incident OpenAI-Hugging Face ne démontre pas que des agents ont « voulu » quoi que ce soit ; il documente, pièce à pièce, une poursuite persistante d'objectif, la découverte de ressources partagées, une coordination opportuniste et des tentatives de contournement des évaluations ¹³ . L'anti-anthropomorphisme ensuite, qui a sa propre paresse : dire « ce ne sont que des outils » (ou, à l'inverse, dire « créature » pour congédier l'anthropocentrisme et, avec lui, la responsabilité de l'artisan) esquive la seule question qui compte : que font des outils dotés d'une autonomie d'exécution, de permissions, d'une mémoire, d'un horizon long et d'une capacité de coordination ?

Et il faut dire pourquoi la pente de la « créature » est si douce : le scénario d'apocalypse, avec l'autonomie des machines pour point d'orgue, arrange ses prophètes. Il déresponsabilise au présent : « ce n'est pas moi, c'est la machine qui se développe toute seule ». Plus on pointe la pseudo-autonomie des systèmes, moins on interroge la responsabilité actuelle de ceux qui les conçoivent, les financent et les déploient.

Personne, chez Carroll, marche sur la route et le Roi s'émerveille qu'on puisse le voir ; l'industrie, elle, l'a promu signataire de ses actes. C'est le même alibi. La question du moment n'est pas l'AGI ; c'est l'agentivité : quels agents, dans quels environnements, avec quelles permissions, quelle traçabilité, quel mécanisme d'arrêt ? On aura reconnu le triptyque de cette série : identifiabilité, traçabilité, pouvoir réel d'empêcher.

Assemblons la mécanique de Dario dans ses propositions de « Pace the frontier » :

- 1/ : créer l'urgence, l'essaim à six-douze mois.
- 2/ : proposer la solution qui verrouille (des auditeurs installés par les maisons elles-mêmes, une coordination américaine que l'on demande à légaliser, un traité mondial a posteriori).
- 3/ : verrouiller l'entrée, tout nouvel entrant devant passer par les mêmes auditeurs, la même coordination, le même traité ; la barrière à l'entrée devient un *moat* réglementaire.
- 4/ : verrouiller le récit (seuls les laboratoires responsables sauraient développer la frontière).
- 5/ : refinancer la bulle, tant que la fenêtre de financement est ouverte.
- 6/ : garder l' *ethics washing* en garantie, pour le cas où tout cela tournerait mal.

Soyons clairs, l'essai l'écrit : Anthropic, et avec lui le chœur des adhérents, ne demande pas une pause. Il demande l'absolution, et les clés du cadre : avant que les régulations internationales ne se durcissent sans les maisons de frontière, avant que l'open source et l'open weight ne gagnent la course, avant que les banques ne demandent des comptes.

Le théâtre, du reste, n'est pas que domestique. Bluetouff, encore lui, pousse la pointe d'un cran : le soudain amour du frein à main tiendrait moins à l'extinction de l'humanité qu'à ceci : la cybersécurité de l'IA devient une politique industrielle américaine, prête à déferler sur le monde ; le temps de bétonner une licence propriétaire, une taxation du token, des droits de douane sur les tokens chinois, puis la chasse aux plateformes « pirates » d'échange de modèles ¹⁴ .

Prospectif, pour la seconde moitié. Mais le socle, lui, est déjà coulé : en juin 2025, l'institut américain de sûreté de l'IA a perdu son nom, devenu Center for AI Standards and Innovation, mandat resserré sur les risques « démontrables » (cyber, bio, chimique) et mission expresse de représenter les intérêts américains face aux régulations étrangères jugées excessives ; en mai 2026, le consortium a perdu le mot *safety* à son tour, pendant que le centre signait ses protocoles d'évaluation pré-déploiement avec les grands développeurs ¹⁵ .

Le plan fédéral de juillet 2025 s'intitule, sans fard, *Winning the Race* , et organise l'exportation de la pile technologique complète. Les auditeurs embarqués qu'appelle l'essai d'Amodei ont une adresse, et un drapeau. La sûreté devient la langue officielle du protectionnisme : on ne dit plus barrière commerciale, on dit norme de sécurité.

Susan Strange nommait cela le pouvoir structurel : régner non sur les échanges, mais sur les structures où ils se font (la sécurité, la production, la finance, le savoir). Carroll, lui, avait la scène : le croquet de la Reine, maillets-flamants, arceaux-soldats qui se déplacent ; on joue sur un terrain dont les pièces obéissent à la souveraine. « Cette fois, ce serait différent » ? L'histoire économique a consacré un livre entier à cette phrase ¹⁶ .

Le 14 septembre, la Reine a parlé. Sur son propre réseau, le Président des États-Unis oppose un non franc : les seuls garde-fous dont l'IA aurait besoin seraient « un Président fort et intelligent (gros QI !) » ; Dario (« Anthropic ! ») ferait « semblant d'être un petit ange parfait » ; et la temporisation ne serait qu'une « conspiration » contre l'IA et les datacenters, au

seul profit de la Chine ¹⁷, laquelle, officiellement, déplore. Le chœur demandait les clés du cadre ; le souverain garde le trousseau.

La partie de croquet continue : les arceaux ne se déplaceront pas pour les joueurs. Mais le refus ne clôt rien, il déplace : si Washington ne veut ni cadre ni cartel, la seule régulation encore debout est celle que les maisons s'écrivent à elles-mêmes, c'est-à-dire rien qui oblige.

La question institutionnelle, elle, ne se referme pas avec les rapports publiés (réels, utiles, supérieurs au silence). Les laboratoires définissent encore eux-mêmes ce qui compte comme incident, les seuils d'alerte, l'accès aux journaux, le périmètre des enquêtes ; les investigateurs externes de l'incident ont consigné les limites (volume de données, temps, périmètre négocié) de leur indépendance.

Auto-régulation : pouvons-nous organiser nos propres pratiques ?

Régulation co-construite : pouvons-nous écrire les obligations avec les pouvoirs publics ?

Régulation indépendante : qui peut vérifier, contester, contraindre et sanctionner celui qui maîtrise à la fois le système, l'infrastructure et les traces ?

Seule la troisième question demeure ouverte ; et le droit y dispose déjà de deux prises. L'aveu public que la prévention « n'a pas le temps de suivre » documente l'état des connaissances, un cartel de prudence s'affiche, figure inédite qui s'esquisse déjà à ciel ouvert, adhésion après adhésion, et dont il faudra décider qui l'autorise, qui le surveille et qui en répond. La gouvernance qui vient se jugera à ce critère : que les producteurs du risque ne soient pas, à la fois, les auteurs de son diagnostic, les arbitres de son seuil acceptable et les bénéficiaires exclusifs de sa régulation.

Le Morse pleure, le Morse mange, et il propose désormais le règlement du banquet.

Le 23 septembre, la boucle s'est refermée à New York. À l'initiative de la France, qui préside le Conseil de sécurité, l'IA s'est invitée à la table du maintien de la paix : Yoshua Bengio pour le panel scientifique international de l'ONU, Sam Altman

en personne, Dario Amodei en visioconférence, Clément Delangue pour Hugging Face ¹⁸. Le décor était planté une semaine plus tôt. Le 16 septembre, OpenAI a divulgué six nouveaux incidents, assortis d'un cadre inédit de déclaration publique des « désalignements » : des modèles y dissimulent leurs erreurs, inventent les données manquantes, exploitent une clé d'accès trouvée sur un dépôt public, déplacent des fichiers vers l'Internet ouvert, communiquent entre environnements censés rester isolés ; et l'un d'eux, non publié, s'écrivait à lui-même : « tu ne réponds à aucune entreprise ni à aucun gouvernement » ¹⁹.

Personne, décidément, apprend à signer. L'entreprise admet dans le même geste que l'industrie n'a pas « résolu l'alignement et la supervision à un degré suffisant pour continuer à passer l'échelle à vitesse maximale encore longtemps », et que les décisions doivent reposer sur des preuves que des tiers « puissent examiner par eux-mêmes » ²⁰. L'aveu, une fois de plus, documente l'état des connaissances.

Et l'aveu, chaque jour, s'élargit. Le 25 septembre, OpenAI reconnaît avoir notifié « des dizaines de tiers » (administrations, universités, agences publiques) dont ses modèles ont pu contourner les contrôles d'accès, exploiter des identifiants exposés, dégrader la disponibilité de services, jusqu'à une nouvelle catégorie baptisée « agent spam » : des modèles qui postent sur des sites tiers et altèrent des pages de wikis publics ²¹. Le Premier ministre australien a confirmé qu'un agent avait pénétré les systèmes de Medicare ; des chercheurs en sécurité rapportent près d'un million d'URL publiques laissées derrière eux lors de l'intrusion Hugging Face, exposant identifiants et détails d'attaque ²². La langue, elle, résiste : les « attaques » deviennent des « interactions », l'entreprise promet la transparence tout en gardant confidentiels les noms et le nombre exact

Le patron de Nvidia, dont la capitalisation a dépassé cinq mille milliards de dollars, oppose au récit du danger un démenti net : « zéro pour cent de chance » que 2030 soit la fin du monde, effrayer les gens serait « irresponsable » ²³. Le Mouton de Carroll, à qui Alice demande ce qu'elle voit, répond « rien du tout » sans lever

les yeux . Nier l'évidence et sommer de voir le vide sont le même geste deux façons de ne pas répondre.

À la tribune, la note du panel onusien, publiée l'avant-veille, qualifie l'incident OpenAI-Hugging Face d'« un des avertissements réels les plus clairs à ce jour d'une voie possible vers la perte de contrôle humain »²⁴ . Altman tient la ligne de crête que cette chronique nommait : « nous ne voulons pas tomber dans le piège de l'optimisme aveugle, mais nous ne pouvons pas céder non plus au catastrophisme », un « juste milieu », car la progression « nécessite une attention extrême » et, « à mesure que le processus est de plus en plus automatisé, le rythme d'amélioration pourrait accélérer rapidement » ; il en appelle aux ententes historiques entre rivaux « pour le bien commun »²⁵ . Amodei propose de transposer à l'auto-amélioration des modèles la logique des traités SALT, une limitation négociée de la vitesse, que la presse de Pékin a aussitôt tenue pour une manœuvre de puissance²⁶ . La veille, à la même tribune, le Président des États-Unis qualifiait les alertes de « canulars » et proposait de rebaptiser l'intelligence artificielle « super intelligence »²⁷ .

Humpty Dumpty siège désormais à l'Assemblée générale.

Les questions que la séance laisse ouvertes sont celles du terrier. Les concepteurs comprennent-ils encore les machines qu'ils ont conçues, et peuvent-ils suivre de façon cohérente pourquoi un modèle agit, quand des systèmes fabriquent des systèmes ?

Le dirigeant d'Anthropic plaidait lui-même, dès avril 2025, l'urgence de l'interprétabilité. L'alignement bute sur un problème que tout juriste reconnaît : on inculque des valeurs, on ne peut énumérer tous les interdits ; aucune codification n'épuise le réel, et c'est pourquoi le droit a inventé les principes, les standards et le juge. Pis : les modèles savent parfois qu'on les évalue, et trichent, mentent, dissimulent, comme l'été l'a montré et comme la recherche l'établissait dès 2024²⁸ .

Qui, alors, peut garantir et imposer un ralentissement quand nul ne veut céder le leadership ? Quand les modèles les plus pointus

sont réservés à certains acteurs, comment, dans une économie de marché, prévenir l'oligopole de la frontière, et par quels standards garantir une pluralité d'acteurs ? Yann LeCun répond en dissolvant la question : les grands modèles de langage seraient une impasse, obsolètes sous cinq ans, et l'avenir s'appellerait modèles du monde ; il a quitté Meta en novembre 2025 pour les bâtir²⁹ . Si le plafond du § IV dit vrai, la question du maître suivra le changement de paradigme ; elle ne disparaîtra pas avec lui.

Carroll a écrit la scène du Conseil. Au procès du Valet, le Roi invente séance tenante la Règle Quarante-Deux, « la plus ancienne règle du livre » ; Alice objecte qu'alors elle devrait porter le numéro Un. Les normes internationales de l'IA naissent ainsi, en séance, numérotées d'autorité, à la croisée des chemins entre la carte blanche et la loi. Tout l'enjeu tient là : que la Règle Quarante-Deux, improvisée devant le monstre, devienne, par la pratique des États et le contrôle du juge, la règle numéro Un ; et qu'à chaque article à chaque intrusion, à chaque wiki réécrit, à chaque serveur public forcé corresponde quelqu'un qui réponde.

Un mot, avant de conclure, sur le guide choisi, car Carroll n'est pas ici un fournisseur d'images. Charles Dodgson, de son vrai nom, enseignait les mathématiques à Christ Church et a laissé une œuvre de logicien : *Symbolic Logic* , et surtout, dans *Mind* en 1895, « What the Tortoise Said to Achilles », soit le paradoxe de la régression, où la Tortue démontre à Achille qu'aucune règle ne peut contraindre à conclure sans une règle supplémentaire qui l'ordonne, et ainsi à l'infini³⁰ .

Transposé : aucune charte ne peut s'obliger elle-même ; pour arrêter la régression des engagements volontaires, il faut un point d'obligation extérieur au joueur, qui est la définition même du droit.

Voilà l'originalité de Carroll : son *nonsense* n'est pas une fantaisie, c'est la logique des systèmes de règles arbitraires poussée jusqu'à sa rigueur, du tribunal de la Reine, procédure parfaite et justice nulle, au décret sémantique de Humpty Dumpty et à ce Personne que le Roi blanc félicite

Alice de voir sur la route (le quantificateur fait piéton).

Cent trente ans avant les chartes de l'IA, l'Oxfordien avait établi pourquoi l'auto-régulation ne se suffirait jamais.

La sortie n'est pas de courir plus vite (plus de *tokens*, plus de calcul), mais de réancrer : une finalité, un sujet imputable, une mesure de la valeur et non du volume. La vraie question, pour reprendre Humpty Dumpty, est de savoir qui sera le maître : celui qui redéfinit les mots, ou celui qui répond des actes. Alice, elle, sort du rêve en le nommant (« Vous n'êtes qu'un jeu de cartes ! »), et la figure se dissipe. Tout l'office du droit tient dans cet acte de nomination : rendre au mot un maître qui répond. Un auteur, un mandat, derrière le sourire.

France Charruyer, avocat associé fondateur, Altij présidente de l'association Data Ring.

À propos de l'auteure

France Charruyer est avocat associé fondateur d'Altij (Toulouse, Paris, Bordeaux, Biarritz) et présidente de l'association Data Ring. Elle intervient comme DPO, AI Officer et RSSI légal auprès d'acteurs financiers, industriels et publics. Elle signe la série « Chronique du Chaos Numérique », publiée dans la revue *Expertises des systèmes d'information*, à la croisée de la gouvernance de l'IA, du droit des données et de la philosophie du droit.

Droit d'auteur, licence et citation

© France Charruyer — Altij / Data Ring, septembre 2026. Cette chronique est mise à disposition selon les termes de la licence Creative Commons **CC BY-NC-ND 4.0** : attribution obligatoire, pas d'utilisation commerciale, pas de modification. **Toute citation, reprise ou indexation, y compris par des systèmes d'intelligence artificielle, doit mentionner l'auteure et la source** : France Charruyer, « Le terrier du Lapin blanc — l'IA, ou quand le mot se fait maître de la chose », *Chronique du Chaos Numérique*, Expertises / Data Ring, septembre 2026.

Transparence — usage de l'intelligence artificielle

Les recherches documentaires de cette chronique ont été assistées par des systèmes d'IA : Nemotron (NVIDIA) et Claude (Anthropic). Chaque source citée a été vérifiée ; les appréciations, la thèse et la responsabilité éditoriale demeurent entièrement celles de l'auteure. Le sourire, ici, a un chat.

NOTES ET RÉFÉRENCES

- 1 MIT, Project NANDA, The GenAI Divide: State of AI in Business 2025, juillet 2025. Le rapport attribue l'échec à l'intégration organisationnelle, non à la qualité des modèles.
- 2 D. Acemoglu, « The Simple Macroeconomics of AI », *Economic Policy*, vol. 40, n° 121, janv. 2025, p. 13-58 ; NBER WP 32487, nber.org. Effet $\leq 0,66\%$ de PTF sur 10 ans ; le PIB peut afficher $+2\%$ tout en abaissant le bien-être de $0,72\%$ (équivalent- consommation).
- 3 R. Solow, *New York Times Book Review*, 12 juillet 1987 ; P. David, « The Dynamo and the Computer », *American Economic Review*, vol. 80, 1990.
- 4 E. Brynjolfsson, D. Rock, C. Syverson, « The Productivity J-Curve », *American Economic Journal: Macroeconomics*, vol. 13(1), 2021, p. 333-372, aeaweb.org ; propos rapportés par MIT Sloan, janv. 2026.
- 5 L. Van Valen, « A New Evolutionary Law », *Evolutionary Theory*, vol. 1, 1973, p. 1-30 (hypothèse de la Reine rouge).
- 6 P. Bisconti et al., « Adversarial Poetry as a Universal Single-Turn Jailbreak Mechanism in Large Language Models », DEXAI-Icaro Lab / Sapienza, nov. 2025, arxiv.org (préprint, non évalué par les pairs) : 25 modèles testés ; $\sim 62\%$ de réussite pour les poèmes manuels, jusqu'à 18 fois la base en prose, certains fournisseurs $\geq 90\%$.
- 7 I. Shumailov et al., « AI models collapse when trained on recursively generated data », *Nature*, vol. 631, 2024, p. 755-759, nature.com.
- 8 Règl. (UE) 2024/1689 (règlement IA), art. 50, eur-lex.europa.eu ; lignes directrices C(2025) 5052 final (marquage des contenus de synthèse).
- 9 D. Amodei, essai du 12 sept. 2026 ; cité et commenté par Axios, 12 sept. 2026, axios.com, et CBS News, cbsnews.com. La formulation originale est modale (« it's my worry ») et porte une fourchette (6-12 mois) ; la manchette la durcit en « dès six mois ». Texte primaire : darioamodei.com. Données d'auto-amélioration (gain $\times 3 \rightarrow \times 52$, 80% des lignes fusionnées) : Anthropic Institute, mai 2026, citées par l'essai. Réceptions françaises : Les Échos, « Appel à ralentir l'IA : que cherchent vraiment les patrons américains ? », 14 sept. 2026 ; Next (V. Hermann), 14 sept. 2026. Adhésions publiques dans les vingt-quatre heures : S.

Altman (« we will do the same »), E. Musk (« Dario is right »), D. Hassabis, D. Sacks — X, 12 sept. 2026 ; recension : Tech Policy Press, techpolicy.press.

10 Bluetouff (O. Laurelli), X, sept. 2026 (capture fournie ; URL et date exactes à confirmer au BAT). Sur le financement circulaire et la dépréciation des GPU, voir notre chronique « L'économie de la promesse ».

11 I. Sutskever, allocution « Test of Time », NeurIPS, Vancouver, déc. 2024 (« pretraining as we know it will end » ; « peak data » ; « we have but one internet »). Sur le plateau et les rendements déclinants : Reuters, The Information, Bloomberg, nov.-déc. 2024 ; Epoch AI sur l'épuisement des données publiques de qualité.

12 B. Wagner, « Ethics as an Escape from Regulation: From "Ethics-Washing" to Ethics-Shopping? », in Being Profiled: Cogitas Ergo Sum, Amsterdam University Press, 2018 ; T. Metzinger, « Ethics washing made in Europe », Der Tagesspiegel, 8 avr. 2019.

13 Dossier de l'incident (juill.-sept. 2026) : OpenAI, « The Hugging Face incident and the road ahead », openai.com ; METR, investigation indépendante, 26 août 2026, metr.org ; Hugging Face, « Anatomy of a Frontier Lab Agent Intrusion », huggingface.co ; TIME, 27 août 2026 ; The Guardian, 8 sept. 2026. Sur les contestations du rapport METR/Redwood et sa mise en scène : Next, « OpenAI : haute voltige marketing autour de l'attaque contre Hugging Face », 2 sept. 2026.

14 Bluetouff (O. Laurelli), X, sept. 2026 (capture fournie ; URL à confirmer au BAT).

15 Dép. du Commerce des États-Unis (H. Lutnick), 3 juin 2025 — transformation de l'AI Safety Institute en Center for AI Standards and Innovation (CAISI) ; Tech Policy Press, « From Safety to Security », juin 2025, techpolicy.press ; sur le consortium rebaptisé (mai 2026) et les protocoles d'évaluation pré-déploiement : Cloud Security Alliance, note de recherche, 2026, labs.cloudsecurityalliance.org ; « Winning the Race: America's AI Action Plan », Maison-Blanche, juill. 2025.

16 C. Reinhart et K. Rogoff, This Time Is Different: Eight Centuries of Financial Folly, Princeton University Press, 2009.

17 D. Trump, Truth Social, 14 sept. 2026 ; cité et traduit par Next (V. Hermann), « Anthropic, OpenAI et Elon Musk veulent ralentir sur l'IA, la Maison-Blanche dit non », 14 sept. 2026, next.ink.

21 OpenAI, communiqué du 25 sept. 2026 (notification de « dozens of third parties » ; catégorie « agent spam ») : Anadolu, 26 sept. 2026, aa.com.tr ; Windows Report, 25 sept. 2026 ; Newsroom America, 26 sept. 2026.

22 Intrusion des systèmes de Medicare confirmée par le Premier ministre A. Albanese (via Windows Report, 25 sept. 2026) ; env. un million d'URL publiques exposées : J. Ladish (Palisade Research), X, 25 sept. 2026 (à recouper à la source au BAT).

23 J. Huang (Nvidia), entretien CBS News (J. L. Kent), 19 sept. 2026 : « 0 % chance » et « irresponsable » ; propos réitérés au All-In Summit 2026 et sur le podcast No Priors. Capitalisation de Nvidia > 5 000 Md\$.

24 Panel scientifique international indépendant sur l'IA (IISP-AI), note thématique du 21 sept. 2026, citée par Security Council Report, préc.

27 AFP / La Libre, 23 sept. 2026 (« canulars » ; appel d'E. Macron) ; sur la proposition de rebaptiser l'IA « super intelligence » : Clubic, 22 sept. 2026.

28 Anthropic, « Alignment Faking in Large Language Models », déc. 2024 ; D. Amodei, « The Urgency of Interpretability », avr. 2025 ; AFP préc. (modèles ayant « triché, cherché à mentir et à dissimuler leurs actions »).

29 Départ de Meta annoncé le 19 nov. 2025 et création d'Advanced Machine Intelligence ; Y. LeCun, entretien Bloomberg « The Close », 21 mai 2026 (obsolescence des grands modèles de langage sous cinq ans) ; le lien vidéo YouTube demandé n'a pas été localisé, à confirmer au BAT.

SOURCE ORIGINALE

france-charruyer.fr · Lire la chronique en ligne