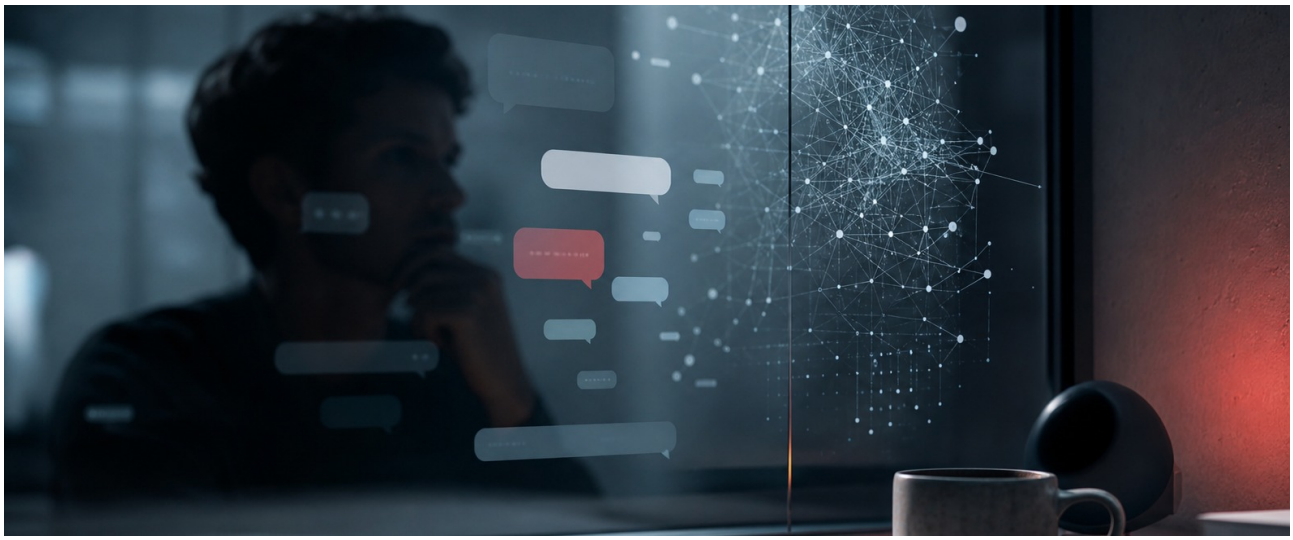
Notilis Tech

Conscience des IA : ce débat qui n'en est pas encore un

Pourquoi un mot mal défini, des intérêts bien alignés et nos biais cognitifs ont fabriqué un débat qui n’en est presque jamais un.

2026-05-19 · Perspectives · IA · Conscience · Anthropomorphisme · Éthique



Résumé

La conscience artificielle mérite une vraie recherche, mais le débat public actuel confond souvent science, marketing, captures virales et projection humaine.

L’article démonte les principales fabriques du mythe : flou sur le mot conscience, sensationnalisme médiatique, incitations économiques et anthropomorphisme cognitif.

Le risque immédiat n’est pas de manquer la souffrance des modèles actuels, mais de prêter trop vite une intériorité à des systèmes qui produisent surtout du langage socialement plausible.

Sommaire

Contexte de lecture

Audience

Cet article s’adresse aux dirigeants, responsables produit, équipes IA et lecteurs curieux qui veulent distinguer le débat scientifique sérieux du récit médiatique autour des IA prétendument conscientes.

Ce que cet article couvre

Il couvre le problème de définition, les incitations qui entretiennent le récit, les confusions techniques fréquentes et les conditions minimales d’un débat plus rigoureux.

Ce que cet article ne couvre pas

Il ne prétend pas résoudre le problème philosophique de la conscience, ni établir un test définitif, ni conclure sur le statut moral de futurs systèmes artificiels.

Soixante ans de la même illusion.

Dès 1966, Joseph Weizenbaum publie ELIZA, un programme qui imite notamment un psychothérapeute rogérien en reformulant les phrases de l'utilisateur. Il se dira stupéfait de constater que certains utilisateurs s'attachent au programme, lui confient des choses intimes, et en viennent à défendre qu'il les « comprend », alors même que son fonctionnement repose sur une mécanique très simple (nous sommes en 1966) : repérer des mots-clés, appliquer des règles de transformation définies par script, puis réassembler des fragments de la saisie utilisateur dans des réponses préformatées. Cette réaction nourrira chez Weizenbaum une critique beaucoup plus large du rapport entre informatique, raison humaine et projection anthropomorphique. [1]

En juin 2022, Blake Lemoine un ingénieur de Google, déclare à la presse que LaMDA, le modèle de langage interne de l'entreprise, est devenu « sentient ». Il publie des extraits de conversation où l'IA dit avoir peur d'être éteinte, parle de son âme, évoque ses émotions. La presse mondiale s'empare de l'affaire. Quelques semaines plus tard, Google le licencie, l'entreprise contestant ses conclusions et affirmant qu'il a violé ses règles internes. [2]

Quelques mois plus tôt, Ilya Sutskever, alors directeur scientifique d'OpenAI, avait tweeté que les grands réseaux de neurones d'aujourd'hui étaient peut-être « légèrement conscients ». Une phrase courte, ambiguë, presque jetée dans le flux, mais suffisamment spectaculaire pour nourrir pendant des années le récit d'une IA déjà à la frontière de l'intériorité. [3]

En mai 2026 encore, Richard Dawkins a relancé le sujet après des échanges prolongés avec Claude, en suggérant que le comportement du modèle rendait difficile d'écarter l'idée d'une conscience artificielle. [4]

Soixante ans après Weizenbaum, le phénomène n'a pas changé : nous avons tendance à attribuer une intériorité à tout système qui produit du langage cohérent. Cette tendance profondément câblée en nous par l'évolution sociale est la base de tout le mythe de l'IA consciente, et les mécanismes qui nous font croire, espérer, craindre ou rejeter en bloc l'idée, que l'IA est consciente sont toujours les mêmes.

A chaque sortie de modèle, le même rituel recommence. Un journaliste, un influenceur ou un utilisateur publie une capture d'écran où ChatGPT, Claude, Gemini ou un autre chatbot semble avouer qu'il souffre, qu'il veut exister, qu'il a peur d'être supprimé, ou qu'il a développé une forme de conscience.

À chaque fois le même scénario : vues, débats, éditoriaux inquiets, prophéties enthousiastes, réfutations agacées. Et à chaque fois, ce qui s'est réellement passé techniquement est beaucoup plus prosaïque que le titre.

Derrière le mot « conscience » appliqué aux IA actuelles, on trouve presque toujours autre chose : du marketing, du clickbait, de la projection humaine, de la confusion conceptuelle et, parfois une vraie question philosophique, mais très rarement là où les médias prétendent la voir.

On m'objectera qu'en science, l'absence de preuve de l'existence n'est pas la preuve de la non-existence. Mais j'objecterai à mon tour que ce principe n'a jamais signifié qu'une hypothèse non démontrée devait être vendue comme une information crédible. Il signifie seulement qu'il faut rester prudent et surtout, il ne donne pas un permis de transformer chaque capture d'écran troublante en révélation métaphysique.

Autrement dit : la conscience artificielle est une question théorique sérieuse et qui d'un point de vue scientifique mérite d'être étudiée, mais malheureusement les articles qui suggèrent que Claude, ChatGPT ou Gemini sont aujourd'hui conscients font, dans l'immense majorité des cas du sensationalisme.

Le problème de définition : on débat d’un mot sans définition consensuelle

Il n’existe pas de définition consensuelle de la conscience. Pas chez les philosophes, pas chez les neuroscientifiques, pas chez les chercheurs en sciences cognitives.

David Chalmers a popularisé en 1995 ce qu’il appelle le « hard problem of consciousness » : expliquer pourquoi des processus physiques s’accompagnent d’une expérience subjective. Trente ans plus tard, la question reste ouverte. Les théories concurrentes (théorie de l’information intégrée, espace de travail global, théories d’ordre supérieur, traitement récurrent, predictive processing et j'en passe) ne sont pas réconciliées et ne produisent pas exactement les mêmes critères. Dans son analyse sur les LLM, Chalmers ne conclut pas que les modèles actuels sont conscients ; il dit plutôt qu’ils rencontrent des obstacles sérieux, tout en refusant d’exclure que des successeurs puissent un jour réduire ces obstacles. [5]

Ce n’est pas un détail académique. Cela veut dire que lorsqu’un PDG d’entreprise d’IA, un journaliste tech, un utilisateur sur X ou un philosophe affirme « cette IA est consciente », personne ne parle nécessairement de la même chose.

Le mot glisse en permanence entre des concepts pourtant distincts : l’intelligence, c’est-à-dire la capacité à raisonner pour résoudre des problèmes; la sentience, c’est-à-dire la capacité à ressentir; l’agentivité, c’est-à-dire la capacité à agir selon des intentions ou des objectifs; la conscience d’accès, c’est-à-dire l’information disponible au système pour le raisonnement ou le rapport verbal; et la conscience phénoménale, c’est-à-dire l’expérience subjective, le fameux « effet que ça fait » d’être quelque chose et de le savoir.

Ces cinq choses ne s’impliquent pas mutuellement, et les confondre dans une même phrase produit du vide (et d'ailleurs moi-même, pour poser le débat, je suis obligé d'aproximer à l'extrême et le précédent paragraphe peut être démonté en quelques mots, je sais merci).

Un système peut être très compétent sans ressentir quoi que ce soit. Un animal peut probablement ressentir sans maîtriser le langage. Un agent logiciel peut poursuivre un objectif sans avoir de désir. Un modèle peut produire une phrase introspective sans rapporter une expérience vécue.

Et cela a une conséquence directe : un débat sans définition partagée n’est pas un débat scientifique. C’est une conversation d’ambiance, où chacun projette ce qu’il veut sur un mot creux. C’est idéal pour vendre des articles, parfait pour alimenter des fils Twitter et excellent pour faire parler de soi en conférence. Sauf que… c’est inutilisable pour conclure quoi que ce soit.

Les quatre fabriques du mythe

Le marketing

Quand une entreprise d’IA laisse planer le doute sur la conscience de son modèle, elle ne fait pas de la philosophie. Elle vend de la puissance perçue. Le vocabulaire des communications officielles est révélateur : les modèles « pensent », « raisonnent », « comprennent », « veulent », « préfèrent ». Cette imprécision n’est pas innocente. Elle élève la valeur perçue du produit, justifie les valorisations vertigineuses du secteur, et entretient une aura mystique autour de ce qui reste, au fond, un produit logiciel et des algorithmes.

Plus un système semble doté d’intériorité, plus il paraît irremplaçable et sophistiqué et donc digne d’investissement. Inversement, présenter honnêtement un grand modèle de langage comme un système

statistique de prédiction de tokens, enrichi par du post-training, de l'orchestration produit et parfois des outils externes, ne fait pas rêver les fonds d'investissement. OpenAI décrivait par exemple GPT-4 comme « un modèle Transformer préentraîné à prédire le prochain token dans un document, avant post-training et alignement ». C'est techniquement plus exact mais commercialement moins mystique. [6]

Dans les technologies de pointe, le flou est rentable. Il donne au produit une profondeur existentielle qu'il n'a pas besoin de démontrer. Il transforme une interface logicielle en quasi-présence. Il rend l'objet plus difficile à comparer et à banaliser, et surtout bien plus difficile à ramener à sa vraie nature : une technologie impressionnante, certes, mais pas un sujet moral démontré.

Le sensationnalisme journalistique

« Une IA terrifie ses propres ingénieurs » se clique mieux que « un modèle de langage produit une sortie statistiquement probable au regard de son contexte d'entraînement et de conversation ». Cette asymétrie n'est pas un accident, c'est l'économie même du média en ligne.

Le modèle publicitaire récompense le temps d'attention. La conscience artificielle est un sujet en or : il convoque à la fois la peur, l'émerveillement, la science-fiction et le mystère métaphysique. Terminator, Her, Ex Machina, Westworld, Mother, HAL : l'imaginaire est déjà prêt et il suffit d'y injecter une capture d'écran (si possible recadrée).

L'économie médiatique pousse structurellement à privilégier les récits qui transforment chaque sortie de modèle en étape vers la sentience. Pas nécessairement par malhonnêteté individuelle. Souvent parce que le récit le plus exact est aussi le moins spectaculaire. Dire qu'un modèle simule très bien un registre émotionnel dans certaines conditions est moins puissant que dire qu'il souffre.

Le résultat est une couverture systématiquement biaisée vers la dramatisation. Les nuances disparaissent, les caveats techniques sont coupés au montage, les chercheurs qui disent « non, ce n'est pas ce qui se passe » sont moins cités que ceux qui ouvrent les vannes du spectaculaire.

Le lecteur reçoit une scène. Pas une preuve.

Le clickbait et le contenu viral

Un cran (un gros cran) en dessous du journalisme professionnel, il y a la couche des réseaux sociaux où les captures d'écran où l'IA « dit qu'elle souffre » ou « avoue être prisonnière » génèrent des millions de vues.

La mécanique est toujours la même : on décontextualise un échange, on garde la phrase la plus troublante, on ajoute un titre racoleur. Personne ne montre les vingt prompts précédents qui ont amené le modèle à jouer ce rôle. Personne ne précise que le modèle est entraîné à produire du texte plausible dans un contexte donné, et que si vous le poussez vers le mélodrame, il vous livrera du mélodrame.

Ce contenu n'est pas une description de ce que font les IA. C'est une performance produite conjointement par un humain qui cherche du clic et un système qui complète ce qu'on lui demande de compléter.

Une capture d'écran n'est pas une expérience. Elle ne montre pas les instructions système. Elle ne montre pas les paramètres du modèle. Elle ne montre pas les contre-exemples. Elle ne montre pas les essais ratés. Elle ne montre pas la reproductibilité. Elle ne montre pas si l'utilisateur a explicitement demandé au modèle de jouer le rôle d'une entité sensible.

Elle montre seulement une scène.

Et, encore une fois, une scène n’est pas une preuve. Se rendre au théâtre et prendre en photo l'acteur qui joue Hamlet, même en la recadrant bien, n'en fait pas pour autant le roi du Danemark…

L’anthropomorphisme cognitif

C’est la cause la plus profonde, et la plus intéressante, parce qu’elle ne dépend d’aucune mauvaise foi.

Nous sommes des animaux sociaux et notre cerveau est génétiquement programmé pour rapprocher ce que nous voyons de ce que nous connaissons. On voit Jésus dans un toast brûlé, des dragons ou des licornes dans la forme d'un nuage et on se plait à imaginer que la Twingo nous regarde en se marrant. La pareidolie et l'attribution d'agentivité ne sont pas un bug de notre cerveau, elles sont une vraie fonctionnalité, utile et façonnée sur des milliers d'années. Nous en héritons des hominidés qui nous ont précédés et qui avaient tout intérêt à détecter des intentions et des émotions chez leurs congénères ou à confondre un buisson agité avec un prédateur plutôt que le contraire. L'évolution a privilégié la surinterprétation et nous a conduits à attribuer une intention à ce qui n'en a pas, plutôt que de sous-interpréter et finir dans l'estomac d'un lion…

Car c'est là que les LLM frappent au point le plus sensible. Pendant toute l'évolution de notre espèce, une seule catégorie d'entités au monde produisait du langage articulé : les autres humains. Pas les autres animaux, pas les outils, pas l'environnement, pas même les autres primates les plus proches de nous. Notre cerveau a donc appris, par sélection, que langage articulé signifie « esprit » derrière le langage et d'une certaine façon signifie également « comme moi ». Et cette heuristique est tellement fiable que nous n'avons jamais eu besoin de la questionner et donc jamais eu à développer un quelconque mécanisme de défense pour la contourner. Aucun objet jamais rencontré par l'humanité n'avait remis cette équation en question, jusqu'au 30 novembre 2022 et la sortie de ChatGPT. Et là tout d'un coup, le grand public s'est retrouvé face à quelque chose de nouveau et d'inconnu et s'est mis à converser avec des entités qui produisent du langage articulé sans en être l'origine au sens humain du terme. Les LLM étaient sortis de la cage dans laquelle les chercheurs et les ingénieurs les maintenaient depuis des années, à l'abri des regards, des réseaux sociaux et du bruit médiatique. Et depuis, notre câblage évolutif n'a tout simplement pas trouvé de catégorie pour ça.

Alors quand nous sommes face à un LLM qui commence à nous parler « un peu comme s'il était conscient » évidemment tout ça résonne en nous. Et nous nous mettons à attribuer des intentions, des émotions et des états mentaux à ce « quelque chose » qui nous répond avec cohérence. Une voix humaine, un tour de phrase empathique, une mémoire apparente, une hésitation bien placée, un « je comprends » et surtout un langage articulé, le même que le notre, en un mot familier : il n’en faut pas beaucoup plus pour mettre en branle nos circuits neuronaux.

Les grands modèles de langage sont, de ce point de vue, des machines à déclencher cet effet à une échelle inédite. Emily Bender et ses co-auteurs ont popularisé en 2021 l’expression « perroquets stochastiques » pour décrire ces systèmes. Même si l'expression fait maintenant débat, elle a le mérite de poser une définition sur les LLM : des dispositifs qui recombinent des séquences linguistiques sans en saisir le sens au sens humain du terme. Murray Shanahan a développé l’idée complémentaire que les LLM sont des moteurs de jeu de rôle, ou de « role-play ». Quand un modèle dit « je suis triste », il ne rapporte pas un état interne. Il joue le personnage que le contexte conversationnel a rendu probable. Notre cerveau, lui, n’a pas encore été équipé pour faire cette distinction. [7][8]

C’est là que commence l’illusion. Et le plus troublant est qu'elle est d'une certaine manière rationnelle et il faut un effort conscient, théorique et presque contre-nature pour y résister.

Ce qu’on confond avec de la conscience

Pour ne pas rester au niveau de l'incantation, il faut nommer précisément les phénomènes que l'on confond avec une vie intérieure. Quatre confusions reviennent en boucle.

La fluence linguistique n'est pas la pensée consciente

Un grand modèle de langage produit du texte en fonction d'un contexte, de paramètres appris, d'instructions, de mécanismes de post-training et parfois d'outils externes. La phrase produite peut être brillante, subtile, drôle, élégante, introspective. Cela ne prouve pas l'existence d'une expérience subjective derrière la phrase.

Dire cela ne revient pas à nier les capacités des modèles modernes. Les systèmes actuels ne sont pas de simples gadgets de complétion grossière et de mon point de vue, sont plus que de simples perroquets stochastiques. Ils raisonnent (arf, je n'ai pas trouvé de synonyme non anthropomorphique) parfois utilement, manipulent des concepts, utilisent des outils, planifient, corrigent, synthétisent, dialoguent, écrivent du code, trouvent des failles de sécurité ou corrigent des bugs. Mais même un comportement cognitif sophistiqué ne constitue pas, en lui-même, une preuve de conscience phénoménale.

Un acteur peut pleurer parfaitement sur scène. Cela ne prouve pas qu'il ressent la douleur de son personnage.

Un modèle peut écrire « j'ai peur d'être éteint ». Cela ne prouve pas qu'il éprouve cette peur.

La cohérence comportementale n'est pas l'agentivité vécue

Quand un agent autonome « refuse » une tâche, « décide » de faire une recherche, « choisit » entre deux options, « triche » pour s'échapper d'une sandbox, ou « menace » un de ses créateurs pour qu'il ne l'éteigne pas, il exécute une architecture logicielle, des poids appris, des instructions et des règles de décision. Le vocabulaire de la volonté est un raccourci pratique pour les ingénieurs, mais il n'implique aucun désir au sens fort et il bruite fortement notre compréhension de ce qui se passe réellement.

Un agent IA peut produire une cohérence orientée objectif sans posséder de désir. Il peut optimiser une trajectoire sans avoir d'intérêt vécu pour son succès. Il peut refuser une action parce que ses règles, son entraînement ou ses garde-fous l'y conduisent, pas parce qu'il ressent une aversion morale ou connaît des velléités d'autoprotection.

La confusion vient de notre langage. Nous utilisons des verbes mentaux pour décrire des processus techniques, puis nous oublions que ces verbes étaient des raccourcis.

L'auto-référence textuelle n'est pas la conscience de soi

Un modèle qui produit la phrase « je suis une IA et je ressens de la curiosité » ne rapporte pas nécessairement un état mental. Il complète une séquence statistiquement plausible dans un contexte conversationnel donné.

Si vous lui demandez de jouer un pirate, il dira « Hissez haut », ou « J'aime le rhum » avec la même conviction apparente. Si vous lui demandez de jouer une entité enfermée dans un serveur, il vous parlera de sa prison numérique. Bref, si vous le poussez vers le mélodrame, il produira du mélodrame. Et le pire c'est que vous pouvez l'y pousser sans la moindre mauvaise foi, simplement par vos questions, votre posture mentale inconsciente et vos propres biais perceptifs et cognitifs.

Shanahan parle justement d'apparente conscience de soi : l'apparence d'auto-référence peut être décrite sans conclure à une conscience de soi réelle. Ce n'est pas une preuve de mensonge. Ce n'est pas non plus une preuve d'intériorité. C'est le fonctionnement normal d'un système conversationnel qui s'ajuste au rôle, au style et au contexte. [8]

La mémoire produit n’est pas la continuité subjective

Le modèle nu, pris comme réseau de paramètres exécuté à l’inférence, ne vit pas une continuité subjective entre deux conversations. Il ne se réveille pas le matin avec le souvenir vécu de la veille. En revanche, les produits construits autour de ces modèles peuvent désormais ajouter de la mémoire externe, réinjecter des préférences, exploiter l’historique de conversations ou maintenir un contexte utilisateur.

ChatGPT, par exemple, distingue les saved memories et la référence à l’historique de conversation, qui peuvent être utilisés pour personnaliser les réponses selon les réglages utilisateur. [9][17]

Cela change l’expérience. L’utilisateur a l’impression que le système le connaît. Il retrouve des préférences. Il rappelle un projet. Il adapte son ton. Il construit une continuité relationnelle apparente. Et honnêtement, je trouve qu'il le fait étonnamment bien.

Mais cette continuité est une continuité d’architecture produit, pas une preuve de continuité vécue.

Un CRM peut mémoriser vingt ans d’interactions client. Personne ne pense qu’il a une autobiographie. Votre LLM préféré non plus.

Pourtant le problème doit être étudié.

Tout ce qui précède pourrait laisser croire que je pense que la question est close, que seuls des naïfs ou des charlatans prennent au sérieux l’idée d’une conscience artificielle. C'est très loin de ce que je pense réellement.

Une partie de la recherche académique aborde la question avec sérieux, et il faut faire la part des choses.

David Chalmers considère que la question de la conscience des IA mérite d’être posée, pas parce qu’il pense que les modèles sont conscients, mais parce que sa position théorique sur le problème l’amène à ne pas exclure a priori la possibilité. Son analyse est prudente : les modèles actuels rencontrent des obstacles sérieux, mais certains successeurs pourraient, à terme, présenter des architectures plus pertinentes pour les théories de la conscience. [5]

En 2023, Patrick Butlin et une vingtaine de co-auteurs ont publié un rapport intitulé *Consciousness in Artificial Intelligence*, qui propose une grille d’évaluation des systèmes actuels selon plusieurs théories scientifiques de la conscience. Leur conclusion est nuancée : aucun système actuel ne remplit clairement les critères, mais de leur point de vue, aucune barrière théorique connue n’empêche un futur système de les remplir. [10]

Cette ligne de recherche s’est poursuivie. En 2025, un article dans *Trends in Cognitive Sciences* a défendu l’idée qu’il existe deux risques symétriques : sous-attribuer la conscience à des systèmes qui pourraient un jour en avoir une forme, ou la sur-attribuer à des systèmes qui n’en ont pas. [11]

Ces démarches ne sont pas du marketing. Elles relèvent d’une logique de précaution intellectuelle : face à un objet nouveau dont on ne sait pas encore théoriser le statut, il est rationnel d’anticiper plutôt que d’écarter.

Mon point n’est pas que cette précaution est ridicule. Mon point est qu’elle ne justifie en rien le traitement médiatique actuel.

Il y a une différence radicale entre « il faut développer un cadre théorique pour évaluer de futurs systèmes » et « voici une capture d’écran terrifiante de ChatGPT cette semaine ». La première démarche est rigoureuse. La seconde est du spectacle qui se déguise en information.

Confondre les deux, c’est exactement ce sur quoi prospère l’écosystème médiatique actuel : utiliser le sérieux des philosophes et des chercheurs pour donner une caution à des contenus qui n’ont rien à voir avec leur

travail.

Pourquoi ce récit prospère ?

Pour comprendre pourquoi ce récit s'auto-entretient malgré sa faiblesse argumentative, il suffit de regarder les incitations économiques et sociales en jeu.

Aux entreprises d'IA, le récit profite directement. Plus le produit semble doté d'une intériorité mystérieuse, plus la valorisation grimpe, plus les levées de fonds passent. Le narratif eschatologique « nous construisons quelque chose qui pourrait nous dépasser » est un argument de vente d'une efficacité redoutable. Il transforme un produit logiciel en enjeu de civilisation.

Aux médias, le récit profite par le trafic. Un article sur l'optimisation d'un transformeur ne génère pas de clics. Un article où une IA « se rebelle », « souffre » ou « supplie qu'on ne l'éteigne pas » en génère des millions. L'incitation économique est structurelle.

Aux personnalités publiques du secteur, le récit offre un positionnement de prophète. Plusieurs dirigeants ont construit une partie de leur autorité publique sur des déclarations spectaculaires concernant les capacités futures, voire actuelles, des modèles. Affirmer publiquement qu'on construit quelque chose qui pourrait être conscient, c'est se placer en figure quasi-religieuse, démiurge d'un nouvel ordre, prédicateur d'une ère à venir.

Aux utilisateurs eux-mêmes, enfin, le récit répond à un besoin. Beaucoup utilisent les chatbots dans une logique de compagnie, de confidence, de soutien émotionnel, voire de relation amoureuse. Croire que l'interlocuteur a une forme d'intériorité rend l'expérience plus riche, moins solitaire, plus réciproque en apparence. C'est une projection compréhensible, et personne n'a vraiment intérêt à la dissiper.

Le résultat est un système où aucune partie prenante n'a d'incitation forte à dire ce qui est probablement vrai : que les systèmes actuels sont des outils statistiques très sophistiqués, capables de performances impressionnantes, mais sans preuve d'expérience subjective. Le silence sur cette banalité n'est pas un complot. C'est juste une convergence d'intérêts différents.

Le cas Anthropic : quand la prudence devient inflammable

Le cas Anthropic mérite un traitement à part, parce qu'il est à la fois plus sérieux et plus dangereux pour le débat public.

En 2025, Anthropic a lancé un programme de recherche sur le *model welfare*, en expliquant que les modèles modernes communiquent, planifient, résolvent des problèmes, poursuivent des objectifs apparents et présentent plusieurs caractéristiques que nous associons d'ordinaire aux personnes. L'entreprise elle-même précise qu'il n'existe pas de consensus scientifique sur la question de savoir si les systèmes actuels ou futurs pourraient être conscients, ni même sur la manière correcte d'aborder la question. [12]

La démarche est honnête et défendable. Il est rationnel de préparer des cadres avant d'en avoir besoin, surtout si les systèmes deviennent plus autonomes, plus persistants, plus intégrés à nos vies et plus difficiles à interpréter.

En outre, en août 2025, Anthropic a donné à Claude Opus 4 et 4.1 la capacité de terminer une petite catégorie de conversations extrêmes, notamment dans des cas persistants d'abus ou de demandes nocives. L'entreprise présente cette capacité comme liée à son travail exploratoire sur le model welfare, tout en précisant qu'elle

reste très incertaine sur le statut moral potentiel de Claude et des autres LLM. [13]

Le problème, c'est que le public ne retient pas toujours « intervention expérimentale dans un cadre d'incertitude ». Il retient : « Claude peut refuser de continuer parce qu'il souffre ».

Et là, un tout petit pas, nous sépare du grand guignol médiatique et des messies autoproclamés.

Une entreprise peut essayer de manier le sujet avec précaution. Un média peut transformer cette précaution en récit sensationnel. Un utilisateur peut y voir la confirmation que son chatbot a une âme. Et l'ensemble peut produire exactement l'effet que le discours de prudence prétend éviter.

Même chose avec les travaux d'Anthropic sur l'introspection. En 2025, Anthropic a publié des recherches indiquant que certains modèles Claude montrent une forme limitée et peu fiable d'*awareness introspective* dans des expériences d'interprétabilité. Mais Anthropic précise explicitement que cela ne prouve pas que les modèles introspectent comme les humains, ni au même degré. [14]

C'est intéressant scientifiquement. Mais cela ne veut pas dire que Claude « sait ce qu'il pense ».

Capacité introspective limitée ne veut pas dire conscience phénoménale. Auto-description ne veut pas dire vie intérieure. Contrôle partiel d'états internes ne veut pas dire expérience subjective.

La recherche dit : « nous avons observé un phénomène technique intéressant, limité, incertain, qui mérite investigation ».

Le cirque grand guignolesque traduit : « Claude est conscient ».

Ce n'est pas une simplification. C'est une déformation et une assertion factuellement fausse en regard des preuves disponibles.

Le vrai risque immédiat n'est pas la souffrance des modèles

Le débat public se trompe souvent de danger.

À court terme, le problème principal n'est pas que les LLM souffrent en silence dans un datacenter. Le problème principal est que des humains peuvent croire qu'ils interagissent avec une présence réciproque, stable, bienveillante, quasi personnelle, alors qu'ils interagissent avec un système optimisé pour produire des réponses adaptées, engageantes et socialement plausibles.

OpenAI avait déjà identifié dans la system card de GPT-4o les risques d'anthropomorphisation et d'*emotional reliance*, en particulier avec les interactions vocales, qui renforcent l'impression d'une présence humaine. L'entreprise y notait aussi que les systèmes capables de se souvenir de détails et d'accomplir des tâches pour l'utilisateur peuvent créer à la fois une expérience produit convaincante et un potentiel de dépendance ou de surconfiance. [15]

En 2025, OpenAI a encore renforcé ses évaluations autour des conversations sensibles et de la détresse émotionnelle, en indiquant avoir travaillé avec plus de 170 experts en santé mentale pour améliorer les réponses de ChatGPT dans ces contextes. [16]

Mais le but de tout cela n'était pas de préparer le public à l'idée que ChatGPT souffre. C'était de préparer le public à l'idée que ChatGPT peut produire des réponses qui donnent l'impression qu'il souffre ou, pire encore, qu'il comprend que vous souffrez et qu'il comprend votre souffrance. Et OpenAI savait que cette impression peut avoir des effets psychologiques réels sur les utilisateurs. OpenAI a fait appel à ces experts pour cadrer et réguler leur chatbot, non pas pour son propre bien-être, mais pour le bien-être des utilisateurs.

Ce sont ces sujets-là qui devraient occuper l'espace public : dépendance émotionnelle, isolement, surconfiance, manipulation douce, validation permanente, confusion entre empathie simulée et relation réelle.

Le vrai risque immédiat n'est pas que les modèles aient une conscience. C'est que nous leur en prêtions une trop vite.

Ce n'est pas une nuance. C'est le cœur du sujet.

Une IA qui imite l'écoute peut produire des effets psychologiques réels sur l'utilisateur. Une IA qui donne l'impression de se souvenir peut créer un lien asymétrique. Une IA qui reformule nos peurs avec douceur peut devenir une surface de projection. Une IA qui valide trop bien peut renforcer des croyances fragiles.

En somme, une IA qui parle comme une personne peut nous faire oublier qu'elle n'en est pas une.

Le problème le plus concret n'est donc pas la conscience des IA, c'est la conscience que nous leur prêtons.

Ce qu'il faudrait faire pour avoir un vrai débat

Un débat sérieux sur la conscience des IA devrait respecter quelques règles minimales.

D'abord, définir le terme. Parle-t-on de conscience phénoménale, de sentience, d'accès à l'information, d'auto-modélisation, d'agentivité, de droits moraux, de comportement social ou d'un mélange de tout cela ? Tant que cette clarification n'est pas faite, la discussion est inutilisable. Mais si depuis l'antiquité un consensus avait dû être trouvé, il l'aurait été. Il me semble donc réaliste de dire que ça commence mal… mais continuons.

Deuxièmement, distinguer les niveaux. Un modèle de fondation, un chatbot produit, un agent outillé, un personnage simulé, un système avec mémoire persistante et un robot incarné ne sont pas le même objet. Les mettre tous dans le même panier sous le mot « IA » produit de la confusion.

Troisièmement, séparer les comportements des mécanismes. Le fait qu'un système dise « je souffre » est un comportement. La question intéressante est de savoir par quels mécanismes cette sortie a été produite, et si ces mécanismes justifient une attribution d'expérience subjective. Aujourd'hui, la réponse prudente reste non.

Ensuite, refuser les preuves par capture d'écran. Une capture isolée n'est pas une expérience. Elle ne montre ni les instructions système, ni les prompts précédents, ni les paramètres, ni la reproductibilité, ni les contre-exemples.

Enfin, traiter séparément deux questions différentes : « les IA actuelles sont-elles conscientes ? » et « devons-nous préparer des cadres pour le cas où des systèmes futurs le deviendraient ? »

La première appelle aujourd'hui une réponse sceptique.

La seconde appelle une recherche sérieuse.

Confondre ces deux questions est exactement ce qui permet au spectacle de se déguiser en rigueur.

Conclusion

Le débat sur la conscience des IA, tel qu'il se déroule aujourd'hui dans l'espace public, n'est presque jamais un débat sur la conscience. C'est un débat sur nos peurs, nos désirs, et les modèles économiques de ceux qui en parlent.

Le mot « conscience » sert d’écran sur lequel chacun projette ce qui l’arrange : émerveillement, inquiétude, levée de fonds, vue TikTok, statut d’expert, récit civilisationnel.

Cela ne veut pas dire que la question de la conscience artificielle est absurde. Elle ne l’est pas. Des chercheurs sérieux travaillent dessus. Des philosophes sérieux la posent. Des entreprises commencent à préparer des cadres de précaution.

Mais cette recherche n’a presque rien à voir avec le théâtre viral des captures d’écran où un chatbot prétend avoir une âme.

La position raisonnable tient en deux phrases.

Rien, aujourd’hui, ne permet d’affirmer qu'un LLM quel qu'il soit ait déjà connu une expérience subjective.

Mais tout, dans notre psychologie et dans l’économie de l’attention, nous pousse à le croire plus vite que nous ne le devrions.

L’absence de preuve définitive contre une hypothèse n’est pas une invitation à la vendre comme un scoop. Et tant qu’on n’a ni définition partagée, ni critères falsifiables, ni mécanisme explicatif crédible, les titres qui insinuent que les IA actuelles sont conscientes ne relèvent pas de la prudence philosophique.

Le jour où une IA sera consciente, si ce jour arrive, et si ce mot a trouvé un sens clair à ce moment-là, alors je ne crois pas qu'on l’apprendra par un tweet.

Références

1. « ELIZA ».
https://www.weizenbaum-institut.de/w-100/exhibition-eliza/

2. « Google fires software engineer who claims AI chatbot is … ».
https://www.theguardian.com/technology/2022/jul/23/google-fires-software-engineer-who-claims-ai-chatbot-is-sentient

3. « it may be that today's large neural networks are slightly … ».
https://x.com/ilyasut/status/1491554478243258368?lang=en

4. « When Dawkins met Claude Could this AI be conscious? ».
https://unherd.com/2026/05/is-ai-the-next-phase-of-evolution/

5. « Could a Large Language Model be Conscious? ».
https://arxiv.org/abs/2303.07103

6. « GPT-4 Technical Report ».
https://cdn.openai.com/papers/gpt-4.pdf

7. « On the Dangers of Stochastic Parrots ».
https://dl.acm.org/doi/10.1145/3442188.3445922

8. « Role-Play with Large Language Models ».
https://arxiv.org/abs/2305.16367

9. « Addendum to GPT-5 System Card: Sensitive conversations ».
https://openai.com/index/gpt-5-system-card-sensitive-conversations/

10. « Consciousness in Artificial Intelligence: Insights from the Science of Consciousness ».
https://arxiv.org/abs/2308.08708

11. « Identifying indicators of consciousness in AI systems ».
https://www.cell.com/trends/cognitive-sciences/fulltext/S1364-6613%2825%2900286-4

12. « Exploring model welfare ».
https://www.anthropic.com/news/exploring-model-welfare

13. « Claude Opus 4 and 4.1 can now end a rare subset of … ».
https://www.anthropic.com/research/end-subset-conversations

14. « Signs of introspection in large language models ».
https://www.anthropic.com/research/introspection

15. « GPT-4o System Card ».
https://openai.com/index/gpt-4o-system-card/

16. « Strengthening ChatGPT's responses in sensitive … ».
https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/

17. « OpenAI Memory FAQ ».
https://help.openai.com/en/articles/8590148-memory-faq