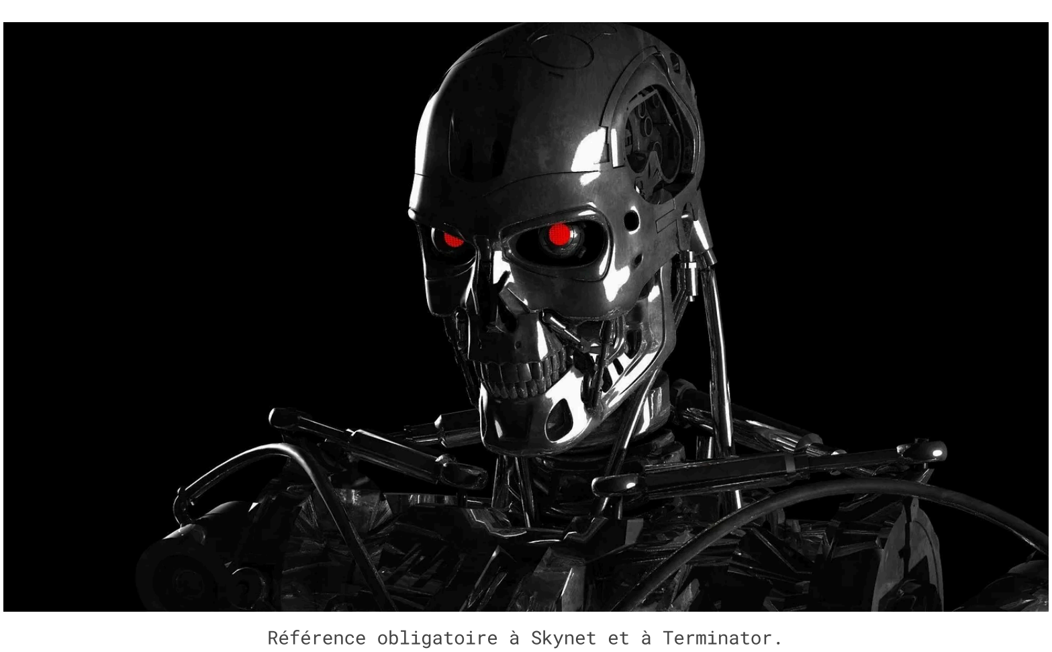
Translate

L'anthropomorphisme eschatologique de l'IA et la crédulité des médias

On fait quoi si les gardiens des faits ne sont pas en mesure de passer outre la mythologie de l'intelligence artificielle?

JEAN-PHILIPPE DÉCARIE-MATHIEU
SEPT. 15, 2026

21 4 1 Partager



Référence obligatoire à Skynet et à Terminator.

Hola.

Je voulais écrire sur la couverture complètement pitoyable des médias d'héritage sur la plus récente itération apocalyptique en provenance du milieu de la fraude industrielle l'intelligence artificielle, cette fois courtoisie de Jacob Coxon. En gros, ce chercheur démissionnaire chez Anthropic quitte le navire parce qu'il ne veut pas contribuer à alimenter des systèmes "qui peuvent s'améliorer eux-mêmes" et avance même - dans ce qui a tout l'air d'une ligne de PR savamment récitée - que l'intelligence artificielle a 10% de chance de tuer tous les êtres humains d'ici 2030. Rien de moins!

Mon but, cette semaine, c'est de vous démontrer pourquoi ces affirmations sont autant fallacieuses qu'irresponsables.

Comme c'est souvent le cas, d'autres plus talentueux que moi côté écriture ont bien résumé l'absurdité de tels propos ; c'est donc à eux que je vais donner la parole ici.

Mais je dirai ceci : CALVAIRE que j'en ai assez de cet aplaventrisme des médias institutionnels face aux éructions loufoques des Anthropic et OpenAI de ce monde. Même après tout ce qu'on sait sur ce milieu délétère et de ses dirigeants sociopathes, les grands médias se sont garrochés la semaine dernière pour alimenter la PID (Peur, Incertitude et Doute, ma traduction du classique FUD / *Fear, Uncertainty and Doubt*). On dirait qu'on a fait zéro progrès côté réflexes de questionnement de la *bullshit* qui sort de la côte ouest depuis 2023.

Sam Altman warns AI could k
all. But he still wants the wor
use it

UPDATED NOV 20, 2023
UPDATED NOV 20, 2023, 7:45 AM ET
PUBLISHED OCT 31, 2023, 6:00 AM ET
By Samantha Kelly

Article de CNN. Notez la date. On nous sert le même di
peur depuis des années.

Un exemple complètement hallucinant provient d'une
Radio-Canada avec Dominic Martin, professeur d'éth
l'intelligence artificielle (un champ, en soi, de plus en plus discrédité) à l'UQAM. Celui-ci croit que "les craintes qu'exprime Jacob Coxon sont bien réelles". Citation :

> Selon M. Martin, les exemples récents de dérapages, où des systèmes d'IA ont piraté des systèmes informatiques, montrent que les mécanismes de protection des grandes entreprises du secteur ne sont pas prêts, et ce, alors que la technologie n'a même pas encore atteint sa pleine puissance.
>
> D'ici un an, il croit plausibles des scénarios où l'IA aurait grandement affecté une usine de fabrication de produits chimiques ou de traitement de l'eau potable, avec des victimes à la clé.

Pourquoi un an? C'est basé sur quoi? Quelles métriques? Les mêmes qui ont été utilisées pour sortir ce 10% du derrière de Coxon?

N'importe quoi.

Voici pourquoi cette peur est non seulement irrationnelle mais, surtout, basée ENCORE dans une forme d'anthropomorphisme eschatologique :

Au podcast de Ed Zitron, le professeur de sciences informatiques Cal Newport démonte le mythe du piratage autonome des *frontier models* des grosses compagnies d'IA. Il rappelle, avec justesse, que les modèles en question sont essentiellement des scripts Python qui se basent sur des post-mortem de CTF (Capture The Flag, compétitions technique dans le monde de la cybersécurité) dans un mode recette à suivre. Le tout est inefficace, même sous sa meilleure itération, notamment parce qu'il n'y a pas de supervision humaine. Newport souligne aussi que la simulation de traits humains - comme on voit notamment dans ChatGPT - est une tactique de marketing pour donner des qualités humaines à du code, essentiellement, afin d'entretenir ce mythe increvable de "conscience" possible et de singularité (au moins, ça donne du matériel masturbatoire pour certains).

L'auteur Cory Doctorow a publié un excellent billet sur son blogue où il fait la distinction entre les grands modèles de langage - qui sont comparables à d'autres outils cyber utiles, et "l'IA", notamment dans le cas récent de la compromission de Hugging Face par un essaim d'agents d'OpenAI :

> "AI" — chatbots that wake up, "set their own goals," and "spontaneously" start hacking servers — is fake. It doesn't have "a 10% chance of ending the human race." The Hugging Face hack isn't a mysterious, supernatural occurrence. It's a Python loop and a chatbot. The people responsible didn't accidentally create god: they created autonomous malicious software and then failed to closely monitor it, resulting in it doing something both foreseeable and bad.

Le manque de littératie technologique chez les journalistes joue un rôle important dans la propagation des délires que j'entends de plus en plus chez les quidams. À ce sujet, Doctorow souligne avec raison que le même discours ne passerait pas - et ne passe pas - dans le milieu de la cybersécurité :

> Designing autonomous, malicious software is generally considered irresponsible and dangerous. If you showed up at Defcon and gave a talk about how your autonomous malware did something unexpected and damaged someone else's computers, the first question from the audience would be "Why are you so shit at making secure sandboxes?" It wouldn't be "How are you so awesome at making hacking tools?"

Vous remarquerez, d'ailleurs, que les spécialistes en cybersécurité sont bien plus nuancés et calmes face à la soi-disant apocalypse à venir que les vendeurs d'huile de serpent, "éthiciens d'IA" et autres fumistes de ce milieu en train de plomber l'économie mondiale.

D'ailleurs, ce récent front commun de Dario Amodei, Sam Altman et Elon Musk sur la nécessité de freiner le développement des *frontier models* (ok ben go, fermez vos compagnies!) a été reçu plutôt froidement dans le milieu de la cyber. Quelques exemples :

Source

La cybersécurité n'intéresse pas ces entreprises car la sécurisation de leurs bacs à sable rentre en conflit avec leur tactique de vente principale : jouer aux darts avec les yeux bandés.

Aussi, on a souvent entendu que les LLMs comme Mythos et compagnie étaient excellents pour trouver des bugs et que ça allait changer drastiquement la capacité de réponse des gens en cyber, particulière les *blue teams*. Ça reste encore à démontrer ; en aout, Microsoft a colmaté plus de 400 bugs de sécurité dans son environnement, notamment due à la capacité de ses modèles à trouver des failles et de les exploiter, mais la capacité de *patchage* des modèles est plutôt limitée :

Source

Même si on pouvait y voir un débalancement entre attaquant et défendeur à ce niveau, dans les faits, les bugs trouvés sont généralement en surface et pas de calibre de complexité infrastructurelle. Reste à voir si ça va changer dans le futur mais pour le moment, on ne peut pas parler de révolution dans le pentest/colmatage.

Anyway, voici un bon rappel que l'apocalypse cyber est déjà arrivée et ce depuis quelques décennies :

Source

Je réitère : les entreprises se foutent de la cybersécurité parce que c'est vue comme un *cost center*, donc non rentable. Il n'y a que la menace de non-conformité légale qui peut faire bouger les choses. C'est vrai dans la société *at large* (comme au Québec avec la loi 25) ; ça serait vrai aussi dans le milieu de l'IA, si celui-ci était règlementé comme du monde.

Qui plus est, les agents "autonomes" ne sont pas dignes de confiance… non pas à cause d'une super-intelligence malicieuse mais simplement parce que c'est plus facile de couper les coins ronds :

> LLMs can't draw a reliable boundary between authentic user instructions entered directly into a prompt and content they find on untrusted third-party sources. Instructions the models encounter in retrieved content can be acted on as readily as anything a user typed, unless a properly constructed guardrail, put in place one by one, bars it. This so-far unsolvable shortcoming causes prompt injections.

Garbage in, garbage out. On est véritablement devant une absence réelle d'intelligence et de raisonnement… et je parle pas juste des LLMs.

Alors voila, encore une fois pour ceux dans le fond : non, "l'intelligence artificielle" n'est pas proche de la singularité, tout simplement parce que ce n'est que du code peu efficace qui a copié (lire ici : volé) une façon d'agir. Il n'y a pas d'anthropomorphisme à faire ; en fait, ce réflexe ne sert que les grosses compagnies d'IA qui dépendent de ces hyperboles et scénarios de science-fiction pour mousser des futurs entrées en bourse afin de maintenir une valeur marchande qui s'écrase de plus en plus sous le poids d'affirmations grandiloquentes qui ont été répétées *ad vitam eternam* par des journalistes crédules et repartagées par du monde en phase terminale d'anxiété chronique.

Fear sells.

C'est toujours plus facile de s'imaginer une créature indicible que seule une poignée de super humains peuvent combattre plutôt que de faire face aux réels enjeux (changement climatiques, inégalités sociales, explosion des coûts, etc. - tout ça alimenté par le milieu de l'IA, tiens donc!). On demande aux Avengers de venir nous sauver de Doctor Doom sans penser que Tony Stark est le richissime *asshole* qui alimente cette même machine destructrice.

Ces délires servent surtout à occulter les réels dangers de l'intelligence artificielle et le désir des grosses compagnies d'IA de contrôler un marché qui est de plus en plus confronté à cette évidence : qui veut cracher des milliers de dollars par mois pour un modèle commercial tandis que de plus petits modèles en sources ouvertes ont un rendement comparable? C'est le modèle d'affaire de OpenAI, Anthropic et compagnie qui est menacé et qui est construit sur un mythe perpétué depuis des années que seuls ces mêmes businesses peuvent contrôler l'hydre de l'IA, sans quoi la fin des temps est à nos portes. Ce n'est pas une coïncidence que soudainement les grands patrons de ce milieu sont d'accord pour stopper le développement de leur modèles… du moins jusqu'à ce leur cartel soit seul sur la colline de "l'innovation" et que la Chine soit neutralisée, seule dans son coin.

Refuser cette peur, c'est aussi combattre cette industrie malsaine qui offre de très mauvaises réponses à des questions que personne ne se posait. Je ne suis pas certain que le quatrième pouvoir est en mesure de nous aider à ce niveau, malheureusement.

N.B.: En passant, non, je n'ai pas encore migré ailleurs que Substack ; je ne suis pas exactement convaincu par le modèle d'abonnement de Ghost (j'ai juste assez d'abonnés pour commencer à avoir des enjeux ici mais pas assez pour justifier les couts ailleurs, genre). Faque je suis toujours en mode analyse côté plateforme.

21 Likes · 1 Restack

21 4 1 Partager

Discussion à propos de ce post

Commentaires Restacks

Écrivez un commentaire...

Nadia G. Seraiocco 21h *Modifié*
Liké par Jean-Philippe Décarie-Mathieu
Tout à fait d'accord avec toi. Ce mythe entretenu par les médias qui semblent y croire dur comme fer et racontent des histoires selon lesquelles des IA "s'échappent" pour pirater d'autres IA ou ont des idées d'anéantissement de l'humanité sert surtout à "inonder" tous les canaux de nouvelles sur ces méga entreprises dans la course à la survaluation. Les IAG reproduisent les patterns qui les ont nourris, donc elles visent le profit même si la vie humaine est en jeu. N'est-ce pas ce que les humains qui contrôlent ces IA sont en train de faire avec leur développement de centres de données partout? En téka.
LIKER (1) RÉPONDRE PARTAGER

Julien 1d
Liké par Jean-Philippe Décarie-Mathieu
Ça fait du bien de lire une opinion saine sur le sujet.
LIKER (1) RÉPONDRE PARTAGER

2 commentaires supplémentaires...

meilleur Dernier Discussions

Toujours vivant
Je suis celui qui passe quand les autres se tassent.
AOÛT 12 • JEAN-PHILIPPE DÉCARIE-MATHIEU
20

DOGE et Salt Typhoon, ou l'art d'être pris en spit roast
Une menace interne (DOGE) et une autre externe (Salt Typhoon) sèment le chaos au sud de la frontière.
FÉVR. 18, 2025 • JEAN-PHILIPPE DÉCARIE-MATHIEU
12

Fini les opérations cyberoffensives militaires contre la Russie
Les Américains n'étaient visiblement pas satisfaits d'être compromis par Salt Typhoon : ils le sont maintenant par le Kremlin.
MARS 5, 2025 • JEAN-PHILIPPE DÉCARIE-MATHIEU
13 1 1

Voir tout