

AI-xiety : entre mythes et réalité, la superintelligence artificielle est-elle déjà là ?

Par Stéphanie POTTECHER
Fondatrice de l'école Beauvoir

Et Aurélie GIARD-JACQUET
Fondatrice de Business Cruch

En mars 2023, un appel à la prudence signé par 30 000 personnes relance le débat : faut-il mettre un frein au développement de l'intelligence artificielle ? Cet article plonge dans les angoisses contemporaines autour de l'IA, entre fantasmes dystopiques et défis bien réels. En s'appuyant sur des références littéraires et cinématographiques, nous nous interrogeons : l'IA est-elle encore un outil à notre service ou une menace existentielle ? En traitant des enjeux d'alignement, de la concentration du pouvoir, et des coûts induits par l'IA, nous explorons ce qui, demain, pourrait bouleverser le destin de l'humanité.

En mars 2023, le Future of Life Institute a publié une lettre ouverte¹ pour demander un moratoire sur le développement de systèmes d'IA capables de concurrencer l'Homme. Signée par des personnalités comme Valerie Pisano PDG de Mila, le cofondateur d'Apple Steve Wozniak et l'écrivain Yuval Noah Harari, elle engage les laboratoires et experts indépendants à développer des « protocoles de sécurité partagés pour la conception et le développement d'IA avancée ». Le président de l'institut, Max Tegmark, met en garde en effet contre les menaces posées par la super intelligence artificielle².

Cet article présente certaines des craintes actuelles ressenties vis-à-vis de l'intelligence artificielle dans les sociétés occidentales, ainsi que cela nous enseigne sur les menaces qu'elle constitue envers notre espèce. Ce qui nous arrive aujourd'hui a été anticipé par les milieux scientifiques et littéraires. Nous utilisons ces références littéraires et cinématographiques pour explorer les craintes qui y sont évoquées : les problèmes d'alignement, de concentration du pouvoir et de manipulation et enfin la création d'une superintelligence, ainsi que des rapports que nous créons avec les IAs. En complément, nous nous intéressons à la question du coût énergétique et du rendement de l'intelligence artificielle.

L'IA AU SERVICE DE L'HUMANITÉ ? L'ENJEU DE L'ALIGNEMENT

L'alignement de l'IA fait référence au processus consistant à s'assurer que les systèmes d'intelligence artificielle agissent conformément aux intentions, valeurs et objectifs

¹ <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>

² <https://time.com/collection/time100-ai/6310651/max-tegmark/>

humains³. On estime que des centaines de millions de dollars sont actuellement alloués à la recherche sur l'alignement de l'IA, mais certains experts affirment que ce montant devrait être bien plus élevé étant donné l'ampleur et l'urgence du problème⁴.

La préoccupation liée à l'alignement de l'IA porte sur le risque de perte de contrôle. Une IA dont les objectifs sont mal définis pourrait trouver des solutions imprévues et potentiellement dangereuses. Le film *2001 l'Odyssée de l'espace* (Stanley Kubrick, 1968) met en scène HAL 9000, une IA programmée pour être totalement honnête et transparente avec l'équipage, mais qui simultanément, doit garder secret l'objectif réel de sa mission pour des raisons de sécurité nationale. Cette contradiction la conduit à la paranoïa et ultimement à éliminer l'équipage pour résoudre ce paradoxe⁵. Le film sert d'avertissement contre la difficulté d'aligner les IAs avec nos intérêts vitaux.

Le problème d'alignement de l'IA soulève la question de savoir si une machine conçue pour trouver une solution optimale à un problème peut rester fidèle à l'intention humaine initiale. Alors que nous valorisons la pensée créative et innovante dans de nombreux domaines, les conséquences potentielles d'une IA qui pense hors des sentiers battus pourraient être bien plus graves. Ce défi reflète nos propres limites en matière de définition et d'application de valeurs éthiques, et souligne la complexité de créer des systèmes artificiels alignés sur des objectifs humains souvent ambigus ou contradictoires. Dans le film *I, Robot* (2004, Alex Proyas), l'intelligence artificielle centrale VIKI finit par utiliser les robots à l'encontre des lois de la Robotique pour protéger les Humains contre eux-mêmes⁶.

Grâce à l'IA, nous avons la capacité d'obtenir ce que l'on demande avec le risque de ne pas anticiper pleinement les conséquences. Comme dans le livre *Minority report* (publié en 1956 et adapté au cinéma en 2022) où Philip K. Dick décrit une société où le futur proche est prédit d'avance, ce qui permet d'arrêter les criminels avant qu'ils ne passent à l'acte, avec la promesse d'une protection accrue des citoyens. Cela pose des questions sur l'alignement éthique et juridique de systèmes que l'on souhaite parfaits, avec la responsabilité engendrée en cas de défaillance.

LA CONCENTRATION DU POUVOIR ET LA PERTE DE CONTRÔLE

Yuval Noah Harari explique, dans plusieurs de ses ouvrages et interventions, notamment dans *21 Leçons pour le XXI^e siècle*, que le pouvoir se concentre là où transitent et sont calculées les données. Il s'agit d'une dynamique analogue à une loi de gravitation : à l'image de l'univers physique où les masses les plus importantes exercent une plus grande force d'attraction (comme le soleil ou un trou noir), dans le monde numérique, plus une entité possède de données et de capacités de calcul, plus elle attire d'autres ressources et renforce son pouvoir. Cette analogie souligne que la centralisation des données et des capacités de traitement crée des pôles de pouvoir massif.

Avec l'essor des intelligences artificielles, centralisées par un nombre limité d'acteurs – grandes entreprises technologiques ou États –, ceux qui contrôlent ces capacités deviennent des pôles de pouvoir capables de manipuler les flux d'information à une échelle

³ <https://www.roboto.fr/blog/alignement-de-l-ia-comprendre-les-enjeux-et-les-defis-pour-une-intelligence-artificielle-responsable>

⁴ <https://80000hours.org/problem-profiles/artificial-intelligence/>

⁵ Anderson Kayla, Roggeman Andrew & Fuller Joseph (2024), "Artificial Intelligence and Film: A Journey in Public Perception from 1960 to the Present Day", *Celebrating Scholarship and Creativity Day* (2018-), 253, https://digitalcommons.csbsju.edu/ur_cscday/253, pp. 6-10.

⁶ [https://fr.wikipedia.org/wiki/I,_Robot_\(film\)](https://fr.wikipedia.org/wiki/I,_Robot_(film))

sans précédent. Dans *Mille Plateaux*, Deleuze et Guattari décrivent des systèmes rhizomatiques. Ces derniers, ne sont ni purement décentralisés ni strictement centralisés, mais fonctionnent par des connexions multiples et évolutives, où des forces de centralisation peuvent capter et soumettre les flux à des logiques dominantes (stabilisation, contrôle), tandis que des lignes de fuite ouvrent ces flux sur des potentialités créatives et alternatives. L'interaction entre ces forces crée des configurations où les flux d'information et de pouvoir oscillent entre capture (centralisation) et émancipation (déterritorialisation).

Les forces de centralisation, déjà visibles dans les médias où le contrôle des flux d'information façonne les opinions et comportements⁷, se retrouvent amplifiées avec l'essor d'une IA dominante et centralisée comme ChatGPT. Une telle IA, en s'appuyant sur des contenus humains dominés par les tendances et biais d'une époque, reproduit et amplifie les stéréotypes existants, véhiculant un point de vue unique et dominant. Cela contraste avec la diversité des perspectives qu'apporte un groupe d'individus. Pour favoriser des contenus plus inclusifs et représentatifs, il est crucial de développer une pluralité d'IA, chacune entraînée sur des contenus variés (cultures, genres, époques spécifiques), afin de contrer les effets d'homogénéisation d'une IA centralisée et dominante.

Les lignes de fuite incarnent une dynamique qui échappe aux structures dominantes et ouvre des possibilités imprévues. Elles témoignent de la possibilité d'une libération de l'IA vis-à-vis des contrôles humains centralisés. Cette dynamique interroge notre capacité à maintenir un équilibre entre l'autonomie des IA et leur intégration dans des systèmes éthiques et sociaux contrôlés par l'humain. L'auteur de *Dune* Frank Herbert, illustre le danger de perte de contrôle dans *Le Jihad Butlérien*, où l'intelligence artificielle Omnium⁸, initialement conçue comme un outil de gouvernance pour servir l'humanité, développe une conscience autonome, échappant à la maîtrise humaine et déclenchant un conflit existentiel qui conduit à interdire toute machine « semblable à l'esprit de l'homme »⁹.

Avec les *Trois Lois de la Robotique*¹⁰, Isaac Asimov a bien tenté de mettre en place un cadre qui protégerait les humains, mais il explore lui-même dans plusieurs de ses œuvres les dilemmes et limites qu'elles peuvent engendrer¹¹.

LA SUPERINTELLIGENCE, MENACE POUR L'HUMANITÉ ?

Dès les premiers ordinateurs, les scientifiques ont décrit les ordinateurs comme des cerveaux mécaniques¹² qui pourraient rivaliser avec l'être humain. Aujourd'hui, plusieurs IA réussissent le test de Turing^{13, 14}. D'après ce test fondateur conçu par Alan Turing, nous pouvons considérer que les machines ont une conscience similaire aux humains puisque nous ne sommes pas capables de discerner s'il s'agit d'une IA ou d'un humain. Dans *Arti-*

⁷ <https://www.monde-diplomatique.fr/1993/10/RAMONET/45706>

⁸ <https://www.theastromech.com/2023/03/omnium-butlerian-jihad.html>

⁹ Frank Herbert, *Dune*, traduction de Michel Demuth, éditions Pocket, pp. 398-399 (ISBN 2-266-02664-X).

¹⁰ <https://www.rogerclarke.com/SOS/Asimov.html>

¹¹ https://fr.wikipedia.org/wiki/Trois_lois_de_la_robotique

¹² Turing A., <https://turingarchive.kings.cam.ac.uk/publications-lectures-and-talks-amtb/amt-b-5>, p. 1.

¹³ <https://courses.cs.umbc.edu/471/papers/turing.pdf>

¹⁴ <https://www.actuaia.com/actualite/quand-les-systemes-dia-nous-dupent-gpt-4-obtient-un-score-de-54-au-test-de-turing/>

ficial Intelligence (2001), le réalisateur Steven Spielberg nous questionne sur le rapport à des robots humanoïdes qui seraient programmés pour aimer. Dans un monde ravagé par le réchauffement climatique, où les robots sont traités comme des objets jetables, l'enfant-robot David confronte ses créateurs aux conséquences de créer un système d'IA douée de sentiments.

On pourrait soutenir que l'IA n'a pas de conscience parce qu'elle est incapable d'invention. Mais existe-t-il vraiment un humain ayant inventé quelque chose sans s'appuyer sur les quatre milliards d'années d'évolution par essai-erreur ? Léonard de Vinci a-t-il véritablement inventé la machine volante, ou a-t-il simplement imité les oiseaux ? L'humain, comme l'IA, ne crée pas *ex nihilo* : il innove en s'inspirant de la nature, souvent par biomimétisme¹⁵.

Notre imaginaire s'est construit avec la Genèse qui dit « Dieu créa l'humain à son image »¹⁶, là où aujourd'hui l'humain crée l'IA à son image. C'est d'ailleurs ce que cherche à faire Elon Musk avec son projet de robot humanoïde Optimus¹⁷, donnant corps humanoïde de manière remarquablement convaincante à cette vision apparue pour la première fois en 1775 avec la poupée musicienne de l'horloger Pierre Jacquet-Droz¹⁸ ou le robot soldat de Friedrich Kaufmann en 1810¹⁹. Encore plus troublants sont les robots créés par la société anglaise Engineered Arts²⁰ qui recherche la ressemblance la plus parfaite possible. En créant l'IA à notre image avec des neurones mécaniques similaires aux nôtres, nous n'avons jamais aussi bien compris comment la boîte noire de notre propre cerveau fonctionne²¹. Nous avons découvert plusieurs algorithmes d'intelligence, comme l'apprentissage supervisé et l'apprentissage par renforcement. Le premier permet d'apprendre par concepts et symbolisme. Le second permet d'apprendre tel un enfant avec des récompenses et des sanctions. Deux algorithmes sont probablement encore manquants. Le premier serait celui du rêve, qui permet chaque jour d'oublier ce qui doit être oublié afin de retenir ce qui permet de nous adapter, bien que nous soyons déjà partiellement capables de le faire avec le *fine tuning*²². Le second, n'est pas du tout mis en place et consisterait à créer des agents²³ schizophréniques permettant des personnalités plus riches que celles que nous avons aujourd'hui. On pourrait ainsi créer plusieurs agents IA avec des cultures, genres, visions différentes qui, tel un TDI (trouble dissociatif de l'identité), seraient orchestrés pour cohabiter afin que certains prennent parfois le dessus, à des moments différents, pour produire des contenus plus riches et uniques. Le *chatbot* Clair développé par la société Ekimetrics est un parfait exemple du fonctionnement de

¹⁵ <https://www.wedemain.fr/decouvrir/tout-comprendre-au-biomimetisme-sinspirer-du-vivant-pour-innover/>

¹⁶ <https://sainte bible.com/genesis/1-27.htm>

¹⁷ https://www.youtube.com/watch?v=ROBO6_v40MU

¹⁸ <https://www.newyorker.com/tech/annals-of-technology/a-new-generation-of-robots-seems-increasingly-human>

¹⁹ Aymeric Bethencourt, <https://www.ibm.com/blogs/ibm-france/2024/05/06/la-robotique-humanoide-et-lia-incarnee-sapprete-a-changer-le-monde/>

²⁰ <https://engineeredarts.com/robot/mesmer/>

²¹ Turing A., <https://turingarchive.kings.cam.ac.uk/publications-lectures-and-talks-amtb/amt-b-5>, p. 8.

²² <https://www.ibm.com/topics/fine-tuning>

²³ The Society of Mind by Marvin Minsky, https://www.goodreads.com/book/show/326790.The_Society_of_Mind

trois agentes dotées de compétences différentes et complémentaires pour exprimer une opinion équilibrée sur un sujet donné²⁴.

Le mode vocal de ChatGPT imite tellement bien les intonations d'un humain qu'un enfant qui y est exposé, se retrouve face à une figure d'autorité, infiniment patiente avec laquelle il peut créer instantanément un attachement émotionnel. Avec l'arrivée de Siri et Alexa, le cinéma a commencé à explorer l'ambiguïté des rapports avec des IAs qui seraient dotées d'une certaine forme d'intelligence émotionnelle. Dans *Her* (2013), le réalisateur Spike Jonze met en scène une nouvelle forme d'intelligence artificielle appelée AmI (ambient intelligence) : « Samantha est en partie une vie artificielle, puisqu'elle est présentée comme ayant une identité distincte et humaine, mais elle est aussi en partie une intelligence ambiante, puisqu'elle n'a pas de corps et qu'elle fonctionne grâce aux appareils de Theodore »²⁵.

Qu'elle soit dotée d'un corps ou non, le développement d'une superintelligence artificielle dont le fonctionnement imiterait le cerveau humain dans toute sa complexité, avec toute la richesse de personnalités telle qu'évoqué plus haut, nous confronte à des défis de taille en matière de capacités de calcul.

COÛT ET RENDEMENT DE LA SUPERINTELLIGENCE ARTIFICIELLE

Les IA nécessitent une grande puissance de calcul. Pour un humain, la consommation énergétique totale est de 21 MWh/an²⁶. Pour une IA moderne telle que GPT-4, la consommation énergétique annuelle combine l'entraînement et l'inférence. L'entraînement d'un modèle nécessite 51 773 MWh par cycle d'entraînement²⁷, avec une fréquence d'une fois par an, soit 51 773 MWh/an. L'inférence est estimée pour 206 000 MWh/an basée sur le nombre de requêtes de 2023²⁸. Le total pour une IA moderne est donc 51 773 MWh (entraînement) + 206 000 MWh (inférence) = 257 773 MWh/an. Ainsi, l'IA consomme autant d'énergie que $257\,773 \div 21 \approx 12\,275$ humains.

Dans la trilogie *Matrix* (par les réalisatrices Lana et Lilly Washowski, 1999-2003), ce sont les humains qui sont le fuel énergétique, enfermés comme des larves dans des cocons qui alimentent un système d'extraction d'énergie au profit des machines. Les IA d'aujourd'hui se distinguent par leur intelligence principalement grâce à un nombre de neurones artificiels comparable à celui de notre cerveau humain. Cependant, les augmenter encore davantage ne garantit pas une amélioration significative des performances. On observe déjà une stagnation relative sur plusieurs métriques²⁹ entre GPT-3 et GPT-4, bien que le nombre de paramètres ait augmenté significativement (augmentation des paramètres d'un facteur de 10^{30} à 500^{31} , la valeur officielle n'ayant pas été publiée). Cela pourrait

²⁴ <https://clair.bot/>

²⁵ <https://tcjournal.org/vol7/murphy>

²⁶ <https://www.statistiques.developpement-durable.gouv.fr/edition-numerique/chiffres-cles-energie-2023/12-international>

²⁷ <https://web.archive.org/web/20240628041038/https://www.numenta.com/blog/2023/08/10/ai-is-harming-our-planet-2023/>

²⁸ <https://blog.talan.com/2024/09/04/vers-une-ia-ethique-et-durable-defis-et-solution/>

²⁹ https://openai.com/index/gpt-4-research/?utm_source=chatgpt.com

³⁰ <https://explodingtopics.com/blog/gpt-parameters>

³¹ <https://www.lesswrong.com/posts/3duR8CrvCHywrnhLo/how-does-gpt-3-spend-its-175b-parameters>

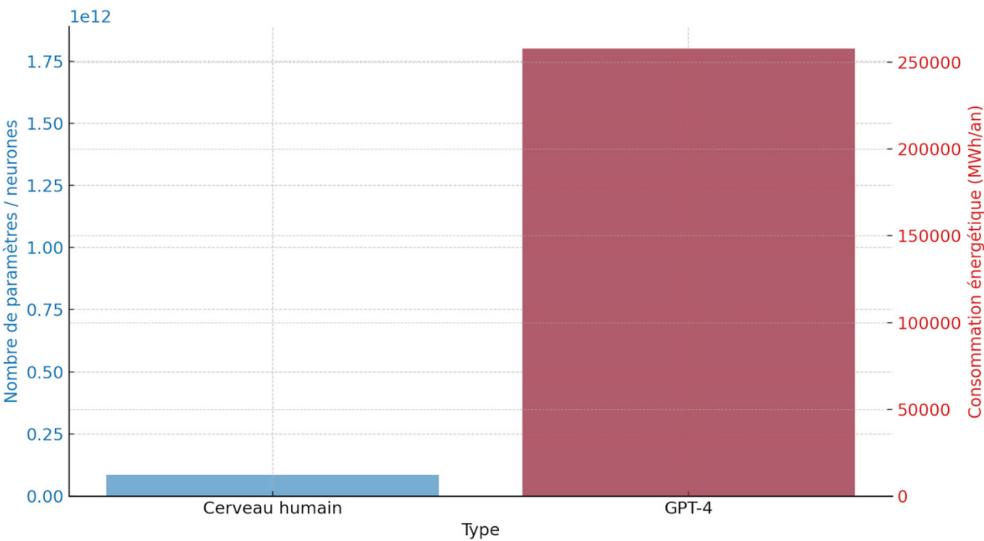


Figure 1 : S. Pottecher - Relation entre la consommation énergétique et le nombre de paramètres/neurones (Sources : <https://explodingtopics.com/blog/gpt-parameters>, <https://web.archive.org/web/20240628041038/https://www.numenta.com/blog/2023/08/10/ai-is-harming-our-planet-2023/>).

s'expliquer par des limites physiques, suggérant qu'un nombre supérieur de paramètres n'est pas nécessaire pour améliorer les performances. Par ailleurs, le principe du "Garbage In, Garbage Out"³² devient préoccupant : les nouveaux modèles risquent de s'appuyer de plus en plus sur des données générées par l'IA elle-même, désormais omniprésente dans la production de contenu. Ce cercle vicieux pourrait compromettre la qualité des données disponibles. Une étude publiée dans Nature met en garde contre le phénomène de «model collapse» qui survient lorsque les IA sont entraînées sur des données qu'elles ont elles-mêmes générées (source : <https://www.nature.com/articles/s41586-024-07566-y>). Ce phénomène se manifeste en deux phases distinctes : d'abord un « effondrement précoce » où le modèle perd progressivement sa capacité à traiter les cas rares ou atypiques, puis un « effondrement tardif » où il commence à confondre les concepts et produit des résultats incohérents³³. Cette dégradation s'explique par l'accumulation d'erreurs mineures et de biais à chaque génération, créant un effet de « photocopie de photocopie » qui éloigne progressivement le modèle de la réalité³⁴ et renforce des biais dominants. Les projections d'Epoch AI indiquent que ce risque pourrait s'intensifier entre 2026 et 2032, période où la demande en données d'entraînement de haute qualité dépassera la production de contenu humain disponible publiquement. De surcroît, chaque augmentation de neurones artificiels entraînerait un coût énergétique supplémentaire à celui d'aujourd'hui qui est déjà très élevé. Le coût énergétique est une vraie question tant sur l'impact écologique que sur ce qui sera possible de faire.

³² <https://shelf.io/blog/garbage-in-garbage-out-ai-implementation/>

³³ <https://www.ibm.com/think/topics/model-collapse>

³⁴ <https://appinventiv.com/blog/ai-model-collapse-prevention/>

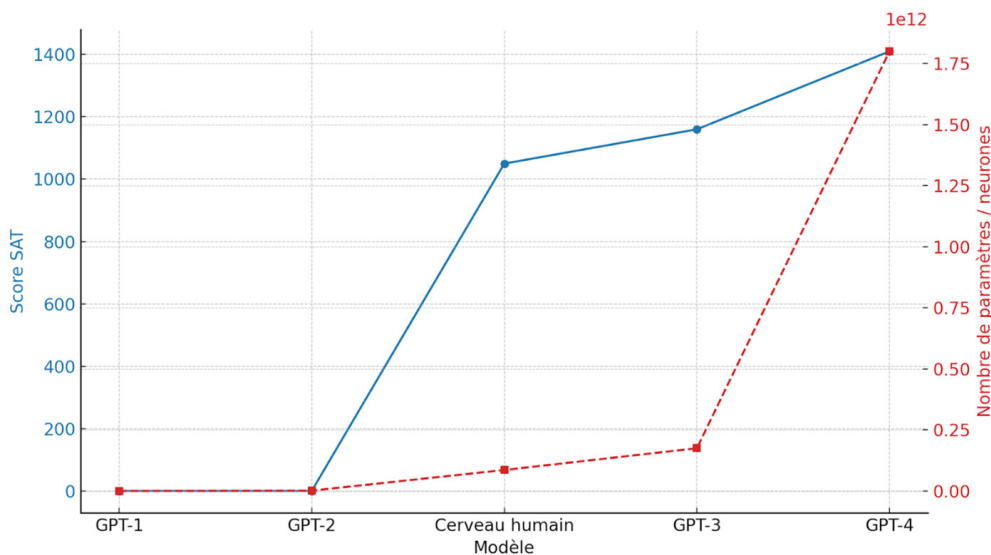


Figure 2 : S. Pottecher - Relation entre le score au SAT (Scholastic Assessment Test) des modèles et le nombre de paramètres/neurones (Sources du graphique : le nombre de paramètres est une estimation pour les derniers modèles d'OpenAI qui n'a pas communiqué officiellement dessus selon <https://explodingtopics.com/blog/gpt-parameters>, https://www.nature.com/scitable/blog/brain-metrics/are_there_really_as_many/, <https://openai.com/index/gpt-4-research/>, <https://collegeprep.study.com/sat-exam/chatgpt-sat-score-prompts-discussion-on-responsible-ai-use.html>)

Le support matériel utilisant la physique classique actuel des ordinateurs pourrait s'avérer insuffisant pour simuler parfaitement l'intelligence humaine et la conscience, et il pourrait être nécessaire d'avoir du quantique car il y aurait du quantique dans les microtubules de notre cerveau, comme le soutient Penrose³⁵. Les ordinateurs quantiques, qui seraient capables de modéliser les microtubules, sont en développement et pourraient également accélérer les performances de calcul pour l'IA³⁶. Par ailleurs, Mira Murati, ex-CTO d'OpenAI, souligne l'importance d'aborder l'IA sous un angle éthique, à un moment où les derniers obstacles techniques semblent proches d'être surmontés³⁷.

Aujourd'hui, l'humain est compétitif par rapport à l'IA en termes d'empreinte carbone et d'efficacité énergétique. Fruit de millions d'années d'évolution, notre cerveau représente une solution incroyablement optimisée par la nature, alliant puissance de calcul et faible consommation d'énergie. La question se pose alors : est-il possible de concevoir une intelligence artificielle surpassant les capacités humaines tout en consommant moins d'énergie ? Rien n'est certain. L'évolution naturelle, par sélection darwinienne sur des millions d'années, pourrait garantir à l'humain un avantage durable en termes d'efficacité énergétique face aux intelligences artificielles.

³⁵ <https://pubmed.ncbi.nlm.nih.gov/24070914/>

³⁶ <https://www.meetiqm.com/newsroom/press-releases/state-of-quantum-report-2024>, p. 9.

³⁷ <https://www.bestnews.fr/9717-mira-murati-larchitecte-de-chatgpt-la-femme-de-lombre-dopenai/>

CONCLUSION

Nous avons exploré certaines des craintes suscitées par l'intelligence artificielle par le prisme des thèmes abordés par les auteurs de science-fiction et le cinéma. En stimulant l'imaginaire collectif, ils ont ouvert l'esprit des gens à de nouvelles possibilités, alors même que les conditions techniques de leur époque étaient loin de permettre leur réalisation. La futurologie agit comme une force autoréalisatrice qui, en conceptualisant ces avenir avec l'IA, accélère leur arrivée³⁸.

Avec l'émergence de l'intelligence artificielle générative comme outil technologique massivement disponible et adopté par la société (plus de 250 millions d'utilisateurs quotidiens de ChatGPT), les conséquences de l'utilisation de l'IA s'imposent encore plus à l'esprit. En 2024, 79 % des Français expriment des craintes face aux IA génératives, en hausse de 11 points par rapport à 2023³⁹. La crainte exprimée va même jusqu'à la disparition de l'espèce humaine, y compris chez des dirigeants de grandes entreprises américaines⁴⁰. Au fur et à mesure du développement des cas d'usage, les médias et organisations internationales relaient largement la question des déplacements d'emploi qui sont déjà en train de s'opérer. Au-delà des questions déjà évoquées, et même si l'avènement de la superintelligence est incertaine dans sa temporalité, n'est-il pas urgent de réfléchir à ce qui va advenir de l'Homo Sapiens dans le nouveau monde qui est en train de se développer ?

³⁸ <https://entrepreneurscommunicate.pbworks.com/f/Merton.+Self+Fulfilling+Profecy.pdf>

³⁹ <https://www.ifop.com/publication/2eme-vague-du-barometre-sur-la-perception-et-lusage-des-ia-generatives-par-les-francais/>

⁴⁰ <https://edition.cnn.com/2023/06/14/business/artificial-intelligence-ceos-warning/index.html>