Traduction de la transcription downsub de la vidéo de Max Tegmark

C'est à ce moment qu'anthropic a commencé à être cotée au Nasdaq, lors de la plus grande introduction en bourse (IPO) de l'histoire. [applaudissements] Maintenant, les sept personnes assises au premier rang que vous avez devant vos yeux nt les sept cofondateurs d'anthropic. Ils valent tous aujourd'hui environ deux ou 20 milliards, pardon. Mais tout le monde ne fête pas ça. Pendant que les dirigeants d'anthropic étaient à Wall Street et dans leurs bureaux de San Francisco, des manifestants se sont rassemblés. Leur demande : qu'anthropic cesse de tenter de parvenir à l'auto-amélioration récursive. Bon, cela peut sembler être un terme très technique. L'auto-amélioration récursive, ou RSI (recursive self-improvement), est le processus par lequel une intelligence artificielle devient si douée en programmation qu'elle peut développer des versions plus intelligentes d'elle-même. Et ces manifestants ne sont pas les seuls à s'inquiéter. Des universitaires et des ingénieurs éminents ont publié une lettre ouverte adressée à l'entreprise. Elle dit ceci. Elle s'appelle la "déclaration misanthropique" et anthropic [Fropic] avertit qu'une ia récursivement auto-améliorée pourrait concentrer le pouvoir et échapper au contrôle humain, alors même que l'on se bat pour la construire en premier. L'entreprise utilise le marketing de la sécurité pour séduire les investisseurs, les employés et le public, alors même qu'elle affaiblit discrètement ses engagements de sécurité et fait pression en faveur de réglementations qui servent ses propres intérêts. Anthropic n'est pas une entreprise d'ia sûre, et nous devons rejeter sa tentative imprudente d'amasser du pouvoir au nom du bien commun.

L'un des signataires de cette lettre ouverte était Max Tegmark, physicien de renommée mondiale et professeur au Massachusetts Institute of Technology (MIT). Il a aussi été en avance sur son époque en matière de commentaires et d'idées sur le développement de l'intelligence artificielle. En 2017, il a publié le livre *Life 3.0 : Être humain à l'ère de l'intelligence artificielle*. Comme vous allez le voir dans cet entretien, ce livre a fait des prédictions qui se sont révélées, je dirais, d'une précision troublante, effrayante même. Et j'ai trouvé cette conversation avec quelqu'un à la pointe de ce débat tout simplement fascinante. Je suis le travail de Max depuis un certain temps maintenant. C'est un gars vraiment, vraiment intéressant. Et surtout, en fait, c'est révélateur, Max explique dans cet entretien ce que les employés d'anthropic admettent après quelques bières. J'espère que vous tirerez de cet entretion autant d'informations que moi.

Max Tegmark, je suis très heureux de vous avoir dans l'émission. Merci de nous rejoindre.

- Merci de m'avoir invité.

- Commençons par parler du sujet de cette lettre ouverte, à savoir l'auto-amélioration récursive et les inquiétudes liées au fait qu'anthropic semble tout à fait disposée à aller de l'avant sur ce terrain. Alors, d'abord, qu'est-ce que l'auto-amélioration récursive, et pourquoi notre public devrait-il s'en inquiéter ?

- L'auto-amélioration récursive, c'est lorsque l'ia elle-même développe la prochaine ia plus intelligente, au lieu que ce soient les humains qui le fassent. Et c'est exactement ce qu'Irving J. Good, dans les années 1960, a soutenu comme thèse, le fait qu'il faudrait aboutir à une singularité et à une explosion d'intelligence. Et d'innombrables personnes, remontant jusqu'à Alan Turing dans les

Voir https://www.youtube.com/watch?v=J9WpNdQYYoo

années 1950, ont dit que cela allait probablement faire perdre aux humains le contrôle de la Terre au profit des machines plus intelligentes qu'ils auront créées.

- Y a-t-il une distinction à faire, parce que concernant l'auto-amélioration récursive, mon idée à ce sujet, ou plutôt celle que j'avais en tête, c'était que la machine se reconstruit en quelque sorte elle-même et devient toujours plus intelligente. Et comme c'est une machine qui le fait et qu'il y a un humain hors de la boucle, il peut y avoir cette sorte d'explosion d'intelligence dont vous venez de parler. Or ce qu'anthropic semble faire ressemble plutôt à des humains qui utilisent simplement des outils d'ia pour se rendre un peu plus efficaces. Y a-t-il donc réellement un modèle d'ia qui s'améliore lui-même actuellement chez anthropic ?

- Excellente question. Tout est une question d'échelle de temps. La civilisation humaine s'est elle-même récursivement auto-améliorée grâce à la technologie. Nous avons utilisé l'imprimerie pour permettre et accélérer le développement d'autres choses. De manière générale, nous utilisons la technologie d'aujourd'hui pour développer celle de demain, et ainsi de suite. C'est pourquoi nous observons une croissance exponentielle. Chaque fois qu'une chose continue de doubler, nous appelons cela une explosion.

Si c'est une bombe nucléaire qui continue de doubler sa production d'énergie, nous l'appelons une explosion nucléaire. Si l'on double encore et encore le nombre d'habitants d'un pays, c'est une explosion démographique. Et si l'on continue de doubler l'intelligence d'une machine... c'est une explosion technologique. Hmm. Maintenant, le doublement de notre économie, par exemple, peut prendre 30 ou 40 ans. C'est donc un processus très lent, et nous, les humains, pouvons piloter en cours de route. Et depuis un moment, il fallait environ un an pour doubler la qualité de l'ia. Nous avons donc beaucoup de temps pour réagir. Mais lorsque vous retirez de plus en plus les humains de la boucle, le doublement se produit de plus en plus vite. C'est ce qu'on appelle la super-exponentielle, en langage de techniciens. Désormais, nous avons de nouvelles améliorations majeures de l'ia qui surviennent à l'échelle de quelques mois, et non plus d'années. Et bientôt ce sera des semaines plutôt que des mois. Et si l'on peut retirer complètement les humains de la boucle, on pourrait imaginer des changements dramatiques se produisant même d'un jour à l'autre. Vous pourriez donc doubler 365 fois en un an, et c'est fou, vous savez.

Si vous prenez un seul atome et que vous doublez le nombre d'atomes que vous avez 365 fois, eh bien, vous avez désormais plus d'atomes qu'il n'en existe dans notre univers. Et c'est vraiment de cela qu'il s'agit. Ma métaphore préférée : si vous prenez votre réacteur nucléaire et commencez à retirer ces barres de graphite qui servent à ralentir les choses, à un moment donné, il va devenir critique, comme on dit, et il va très, très rapidement exploser.

- Et à quoi cela ressemblerait-il ? Je veux dire, à quoi ressemblerait une explosion d'intelligence ? Comment saurions-nous qu'elle s'est produite, disons ?

- On voit beaucoup de données indiquant que ça commence à se produire. Donc anthropic se targue désormais d'environ 80 % de leur programmation réalisée par l'ia, et il reste encore 20 % d'humain, ce qui ralentit les choses. Mais regardez plus largement : je dirais que des années 1950 jusqu'à environ 4 ans en arrière, l'ia a été chroniquement survendue, promettant toujours trop, livrant trop

peu. Et depuis, elle a été très sous-vendue. Il y a seulement quatre ans, les machines sont enfin devenues capables de tenir une conversation raisonnable avec chatgpt. Il y a deux ans, elles se trompaient encore sur des problèmes de mathématiques élémentaires. L'an dernier, elles ont obtenu la médaille d'or à l'Olympiade internationale de mathématiques. Cette année, en écrasant tellement de problèmes de mathématiques célèbres que beaucoup de gens commencent à abandonner leur doctorat en mathématiques, se sentant sans avenir dans cette discipline. Et l'ia est soudainement capable de piratage surhumain, et il y a aussi des essaims de drones agressifs qui se déchaînent et font des choses que nous, humains, ne pensaient pas qu'ils puissent faire. Donc on voit dans les données une très nette accélération, une montée en vitesse du rythme des progrès. Et ce qu'Alan Turing, le parrain de tout notre domaine, a dit en 1951, c'est que si nous fabriquons des machines extrêmement plus intelligentes que nous, nous devons nous attendre à ce qu'elles prennent le contrôle.

C'est assez évident d'une certaine manière, car nous ne comprenons pas très profondément comment fonctionnent les meilleurs systèmes d'ia actuels. C'est embarrassant pour moi de le dire, en tant que professeur qui fait de la recherche en ia depuis une décennie ici au MIT, nous avons beaucoup travaillé sur ce domaine. Cela s'appelle l'interprétabilité mécaniste : essayer de comprendre comment ils fonctionnent. Nous ne comprenons pas comment, en particulier, ils fonctionnent, ni quels objectifs ils ont, ni même comment mesurer quels objectifs ils ont. Donc si maintenant vous imaginez que vous avez des millions, des milliards ou des milliers de milliards d'agents ia actifs dans le monde, et que nous leur donnons des corps robotiques comme Elon Musk et d'autres veulent le faire, qu'ils font leurs propres choses... il est assez difficile de voir un scénario où nous pourrions rivaliser efficacement contre eux, où nous pourrions rester pertinents et garder le contrôle sur eux.

- Donc vous avez dit là quelque chose : s'ils reçoivent tous des corps robotiques comme Elon Musk le veut. Et je pense qu'il serait probablement vrai que s'ils recevaient tous des corps robotiques, les gens seraient bien plus effrayés qu'ils ne le sont actuellement. Pourquoi les gens ne devraient-ils pas attendre que les robots soient produits ?

Parce que si au cours des 12 à 18 prochains mois - ce qui semble préoccuper beaucoup de monde - anthropic [Fropik] parvient à une explosion d'intelligence, tout cela se fera en termes de logiciel. Cela n'aura pas vraiment de réalité physique, vraiment. Dès que vous avez beaucoup d'argent et beaucoup d'intelligence, même sans corps, vous pouvez payer des gens pour travailler pour vous. Vous pouvez organiser des entretiens d'embauche par zoom. Vous pouvez même faire passer ces entretiens à votre avatar, si réaliste que les gens croiront travailler pour un humain. Vous pouvez engager des cabinets d'avocats pour vous défendre. Vous pouvez engager des gardes du corps, des mercenaires. Nous voyons déjà l'ia faire un peu de cela. Comme celui qui s'est échappé d'openai et a tenté de convaincre un mainteneur turc d'un certain logiciel sur le réseau qu'il était un humain et qu'il devait accepter du code malveillant qu'il avait mis en ligne. Quand vous avez de l'argent, il est assez facile de le convertir en troupes au sol, en gens qui font des choses pour vous dans le monde réel.

- Pensez-vous que votre rôle, votre métier - je veux dire, vous êtes un expert de renommée mondiale en physique et en cosmologie -, l'une des disciplines, je suppose, qui, par son intersection avec les mathématiques, est l'un des domaines où l'ia avancée progresse le plus spectaculairement, pensez-vous être dans une position particulière où vous êtes très conscient des progrès accomplis dans l'ia, potentiellement à un degré que beaucoup d'autres personnes de professions différentes ne le sont ?

- Oui, je le pense. Parce que je consacre une fraction significative de mon temps à la recherche sur des sujets liés à la physique et aux mathématiques où je me contente de me dire : “Oh, comme c’est cool, l’ia sait maintenant parler et faire de belles vidéos”. Je fais des choses où j’ai le sentiment de me dire : “Oh mon Dieu, cette analyse physique que cette chose m’a donnée est en réalité meilleure que tout ce que j’ai jamais entendu de la part d’un quelconque professeur ici au MIT”. Vous savez, je pense que c’est l’un des aspects insidieux de cette histoire. Il était facile, il y a quatre ans, pour tout le monde de remarquer que l’ia était passée de l’incapacité à parler à la capacité de parler. Mais pour la plupart des gens, il est bien plus difficile de savoir quand l’ia devient, disons, bien meilleure en mathématiques que leur professeur de lycée, par opposition à bien meilleure en mathématiques que le meilleur mathématicien du monde. C’est un peu comme pour moi : franchement, je ne saurais faire la différence entre, disons, un professeur de violoncelle de lycée et Yo-Yo Ma, parce que je n’ai simplement pas passé assez de temps au violoncelle pour être capable d’apprécier ces nuances. Donc je pense que c’est pourquoi la plupart des gens ne réalisent pas à quel point les changements récents ont été radicaux et rapides.

- Je veux dire, vous êtes professeur au MIT en physique et cosmologie, il n’y a vraiment pas beaucoup de gens comme vous. Donc ce que vous êtes en train de dire actuellement, c’est qu’il y a potentiellement un développement inquiétant : seule une très petite portion de la société peut réellement percevoir l’étendue de l’avancée de l’ia.

- Je lis des articles à ce propos, mais je ne l’ai pas vue, vous savez, je discute simplement avec l’ia. Je ne l’utilise pas pour des expériences de physique avancées, donc je ne peux pas ressentir son intelligence de la même manière que vous.

- Les personnes qui la ressentent le plus fortement, je dirais qu’en ce moment ce sont celles de la région de la Baie, principalement à San Francisco et alentour, qui développent ces systèmes, parce qu’ils passent toute la journée avec elles et remarquent que l’ia devient meilleure qu’eux dans leur propre travail. Beaucoup de gens chez anthropic s’attendent à être licenciés dans un an environ, lorsque l’ia les rendra redondants, par exemple. Et je trouve assez choquant que lorsque je rentre de là-bas - où tout le monde est tellement terrorisé et pense que l’année prochaine pourrait être la singularité - pour revenir ici, dans le Massachusetts, où c’est comme si rien ne se passait. Je pense que cela est en train de préparer la plus grande faillite d’intelligence [échec du renseignement] aux États-Unis depuis le 11 septembre, parce que le gouvernement et nos services de renseignement ne la ressentent pas, eux non plus, au niveau où ils le devraient. Nous voyons maintenant, pratiquement chaque jour, une autre panne massive dont les membres de notre gouvernement sont très rapides à dire : “Rien à voir ici, ce n’est probablement pas lié à l’ia”. Et ce n’est bien sûr que le début de quelque chose. Je prédis, retenez mes paroles, que cela va empirer, et encore empirer, très bientôt. Je pense que la véritable grande erreur d’intelligence ici, c’est de ne pas réaliser que c’est en réalité une urgence nationale, et c’est comme, vous savez, le fait que les gens soient très conscients qu’il existe d’autres pays pouvant représenter une menace massive pour l’Amérique. Mais je pense que d’ici la fin 2027, ces personnes réaliseront que la plus grande menace est en réalité de ce pays-même, de génies dans un centre de données dont Dario (Almodei) dit que cela va advenir l’année prochaine. Vous savez, avec un nombre énorme de ces agents ia qui sont tout simplement meilleurs que n’importe quel prix Nobel en tout, et qui ne sont pas sous le contrôle du gouvernement américain et ne

partagent pas nécessairement les objectifs du gouvernement américain. Donc je pense qu'il est en réalité utile de considérer qu'une troisième superpuissance est en train de s'élever ici. Je pense que c'est une menace à la fois pour l'Amérique et pour la Chine, ce qui est, je crois, l'une des raisons pour lesquelles Xi Jinping et Donald Trump ont désormais des conversations constructives l'un avec l'autre.

- Et parlons des prédictions, parce qu'évidemment, vous êtes un professeur prestigieux, mais il y aura beaucoup de gens qui écouteront cela en se disant : "Tout ça a l'air un peu farfelu". Qu'est-ce qui se passe ici ? Dans ce livre, écrit il y a 10 ans, donc avant la fondation d'openai, bien avant celle d'anthropic, dans le premier chapitre, vous commencez par une histoire illustrative qui parle d'un groupe de personnes dans une entreprise d'ia qui sont déterminées à construire une intelligence artificielle générale afin de pouvoir, je suppose, répandre leurs valeurs dans le monde. Et elles ont des valeurs assez particulières, et elles veulent y parvenir en pratiquant l'auto-amélioration récursive, et elles pensent que même si c'est très dangereux, la raison pour laquelle elles doivent le faire, c'est que seules elles peuvent être dignes de confiance pour contrôler l'ia générale, la super-intelligence, lorsqu'elle sera construite. Et le programme qu'elles construisaient dans votre exemple s'appelait Mythos, n'est-ce pas ?

- Non, il s'appelait Prometheus.

- Pardon. Prometheus.

- Mais maintenant, c'est hilarant à un tel point... Car maintenant nous avons mythos créé par un autre petit groupe de gens versés en technologie, avec des valeurs assez singulières qu'ils veulent imposer au monde, et qui nous disent que nous devons leur faire confiance pour courir vers l'intelligence artificielle générale via l'auto-amélioration récursive, parce que seuls eux peuvent être dignes de confiance.

- Est-ce une prédiction, ou pensez-vous que Dario Amodei a simplement lu votre livre et s'est dit : "Oh, en fait, ça a l'air d'une excellente idée, faisons-le ?"

- [*Rires*] Vous savez, je voudrais vraiment avoir eu tort pour les prédictions de ce livre. Je ne comprends pas pourquoi les gens veulent avoir raison. C'était amusant. Demis Hassabis, qui a fondé google deepmind là-bas, au Royaume-Uni, et en a été le PDG pendant de nombreuses années, a dit, après la sortie de ce livre, lorsque j'ai discuté avec lui dans son bureau : "Eh bien, vous savez, ce n'était pas exactement ainsi que je comptais procéder", puis il a ri. Et je pense que c'est effectivement ce qui est vrai, parce qu'il était très axé sur la construction d'outils d'ia contrôlables, aidant la science, etc., mais il n'a jamais voulu construire des êtres, des êtres ia. Il voulait construire des outils.

Anthropic, en revanche, oui, a un comportement qui ressemble beaucoup à celui du système de l'histoire de mon livre, dans la mesure où la société s'appuie réellement sur le transhumanisme et ils ont cette idée de vouloir créer une ia consciente. Chez anthropic, on parle de droits des machines, vous savez. Et quand je parle à des gens qui travaillent chez anthropic, après avoir bu quelques verres, la conversation aboutit souvent à l'idée de renverser le gouvernement américain et de prendre le contrôle du monde. Et je pense que pendant un moment, je n'arrivais pas à comprendre pourquoi

davantage de gens ne réagissaient pas à leurs discours sur la prise de contrôle du monde, mais ensuite j'ai compris pourquoi. C'est parce qu'ils parlent en code. Ils ont tendance à ne pas utiliser de expressions comme "coup d'état" ou "renverser le gouvernement américain". Ils utilisent à la place la formule codée : "alignement sans contrôle". Alors laissez-moi traduire ce que cela signifie, vous savez. S'il s'agissait de contrôle, n'est-ce pas, s'ils avaient un plan pour que cela soit réellement contrôlé, alors le gouvernement américain pourrait le contrôler. Il pourrait l'arrêter, il pourrait avoir des interrupteurs d'arrêt d'urgence, etc. Mais ils ne veulent pas de cela. Ils ont pour ainsi dire cessé d'en parler. Au lieu de cela, ils parlent d'alignement, ce qui, très grossièrement parlant, signifie que l'ia devrait être gentille, et aller vers certaines valeurs, qui seront probablement assez influencées par les valeurs particulières qu'ils ont, mais sans être contrôlée. Ce qui signifie que nous obtiendrons la sorte d'avenir qu'Alan Turing avait prédit, où ce sera le dieu-machine d'anthropic qui gouvernera la Terre.

Et le reste d'entre nous n'aura qu'à espérer qu'il sera d'une certaine manière gentil. Qualifiez-moi de démodé, si vous voulez, mais j'ai un faible pour la démocratie. L'idée que si l'on veut réellement renverser un ensemble d'institutions démocratiques, on devrait organiser un référendum à ce sujet. Mais ces entreprises ne semblent vraiment pas aimer ça. Elles ont fait pression très durement, ici aux États-Unis, pour empêcher toute surveillance gouvernementale significative sur elles. Anthropic et openai sont moins réglementées qu'une boutique de sandwichs. Elles ont en fait le droit légal de libérer une super-intelligence demain, si jamais elles disposaient d'une telle super-intelligence et si elles le voulaient.

- Restons un instant sur ce moment, parce que, vous savez, moi-même, nos téléspectateurs, il n'y a pas beaucoup de gens dans le monde qui ont l'occasion de boire un verre avec les gens d'anthropic et de leur parler quand ils sont plus décontractés, quand ils ne sont pas aussi contrôlés qu'en public. Et ce que vous dites, c'est que vous avez bu avec des gens chez anthropic occupant des fonctions importantes là-bas, qui, après quelques bières, parlent ouvertement de renverser le gouvernement, et d'avoir claude ou Dario Amodei... Je ne sais pas lequel c'est. C'est le robot ou l'humain contrôlant le robot qui gouverne tout ? Bref.

- Claude. C'est claude.

- Ils veulent, et ils croient sérieusement qu'il est possible que claude gouverne tout.

- Oui. Et soyons clairs, je ne suis pas en train de vous révéler quelque grand secret qu'ils ne disent pas aussi publiquement. C'est juste que quand ils parlent publiquement, ils parlent toujours en code et disent : "Oh, l'alignement, l'alignement". Nous allons rendre claude aligné. Non, non, non, pas contrôlé. Ça ne serait pas gentil envers lui et bla bla bla, mais l'alignement. Et je répète ce que je disais : c'est exactement ce que cela signifie. Ils ne veulent pas que les humains restent aux commandes. Ils n'ont même pas de plan pour cela. Bien au contraire, ce qu'ils essaient si fort de faire, c'est de résoudre l'alignement et de trouver un moyen pour que l'ia soit vraiment gentil une fois que nous aurons perdu son contrôle.

- C'est tellement similaire, n'est-ce pas, à l'introduction de votre livre où vous présentez ce scénario, parce que ce sont des gens assez bien intentionnés dans votre livre, n'est-ce pas ? Ils veulent

mettre fin à toutes les guerres, non ? Ils veulent mettre fin à tous les conflits. Ils veulent que tout le monde vive dans une société plus égalitaire. Évidemment, ils seront eux-mêmes très riches, mais vous savez, ils veulent que ce soit une utopie pour nous.

- le futur, maintenant que presque tous sont devenus millionnaires et, après que nous ayons évoqué hier... Pensez-vous que l'argent affecte leur façon de penser ? Parce que je suis sûr que vous connaissez Dario Amodei, je ne sais pas si vous êtes amis, mais je pense que vous vous êtes rencontrés dans le passé, n'est-ce pas ?

- Il est venu ici, chez moi, juste ici.

- Et parce qu'il a quitté openai parce qu'il pensait que Sam Altman n'était pas assez sérieux au sujet de la sécurité, ou c'est du moins ce qu'il a dit être la raison de son départ.

- Oui.

- Et puis il a déclaré publiquement, à de nombreuses reprises, que l'auto-amélioration récursive est tellement dangereuse qu'il faudrait peut-être ne pas la pratiquer. Et maintenant, dans leur prospectus, dans leur IPO, ils disent littéralement aux investisseurs que la raison d'investir dans anthropic, c'est parce qu'ils sont les plus proches de l'auto-amélioration récursive. A-t-il été aveuglé par l'argent ? Est-ce qu'il mentait depuis le début ? Qu'est-ce qui se passe ici ?

- Upton Sinclair a célèbrement lancé qu'il est difficile de faire comprendre quelque chose à un homme lorsque son revenu dépend du fait de ne pas le comprendre. Donc je ne pense pas que Dario, ou aucun des autres PDG d'ailleurs, mente au sens où ils diraient des choses qu'eux-mêmes savent fausses. Je pense qu'ils se mentent à eux-mêmes, toutefois. Le chemin de l'enfer est pavé de bonnes intentions, n'est-ce pas ? À peu près tous ceux qui ont causé de grands dommages au fil de l'histoire se sont convaincus qu'ils étaient les gentils.

- J'aimerais revenir sur certains des aspects techniques de la question, parce que je suppose qu'il y aura toujours des gens qui, de façon tout à fait compréhensible... Et je veux dire, je ne suis pas certain à ce sujet, vous savez, je pense qu'il est tout à fait plausible que même si ces gars veulent faire cela, il y aura un obstacle qui rendra la chose impossible, ou que nous avons peut-être mal conceptualisé l'ia super avancée et que cela ne se passera pas du tout comme ça. Alors passons en revue certains de ces obstacles possibles. Et l'un d'eux concerne la physicalité des centres de données, parce que d'après ce que je comprends, l'auto-amélioration récursive, l'idée, c'est que vous créez essentiellement de très bons codeurs pour qu'ils puissent améliorer les algorithmes impliqués dans l'apprentissage automatique. Mais l'apprentissage automatique requiert aussi une sorte de nombre exponentiellement croissant de GPU, de puces fonctionnelles.

- Puis-je vous interrompre juste là : aujourd'hui, c'est le cas, mais je pense que très bientôt ce ne le sera plus. Je pense qu'il est assez clair qu'on n'a pas besoin d'un gigawatt pour créer quelque chose doté d'une intelligence de niveau humain, parce que votre cerveau fonctionne en ce moment-même avec 20 watts.

Et l'un des aspects essentiels de l'auto-amélioration récursive, c'est que l'ia trouve comment modifier radicalement ses algorithmes. Ainsi, avec la même quantité de calcul, elle peut être simplement infiniment plus intelligente, d'accord ? Je pense que jusqu'à présent, nous avons essentiellement utilisé des algorithmes vraiment stupides, et que nous avons contourné le problème en y jetant une force brute, avec bien plus de matériel que nécessaire. Et nous nous dirigeons bientôt vers une époque où nous n'aurons pas besoin d'énormes quantités de données. Je veux dire, je pense que nous en avons déjà largement assez, des centres de données, pour, avec de meilleurs algorithmes, créer une ia capable de renverser le gouvernement américain et de prendre le contrôle du monde. Donc je ne vois pas cela comme un obstacle. Je pense que jusqu'à il y a environ deux ans, beaucoup de gens espéraient qu'il existait un autre obstacle technique qui nous empêcherait d'atteindre le stade de l'auto-amélioration récursive. En réalité, il y a 6 ans, presque tous les chercheurs en ia que je connaissais pensaient que nous étions à des décennies d'une ia capable ne serait-ce que de passer le test de Turing, c'est-à-dire de maîtriser le langage et les connaissances approximativement au niveau humain. Et il a suffi d'environ 3 ans pour que cela se produise effectivement. Et maintenant, je pense, en parlant maintenant en tant que chercheur en ia moi-même, qu'il y a très peu d'indices suggérant que nous ne puissions simplement avancer de force avec les techniques existantes et parvenir à l'auto-amélioration récursive. Je pense qu'il existe un autre obstacle énorme. Donc si nous parvenons à un avenir vraiment bon, où nous profitons tous d'outils d'ia puissants et contrôlables dont les humains ont la maîtrise, capables de guérir le cancer et de nous rendre prospères, etc., ce sera grâce à cet obstacle-là, qui sera l'obstacle humain. Il y a ce vaste mouvement pro-humaniste qui se lève actuellement ici aux États-Unis : la coalition de Bernie à Bannon, avec Ron DeSantis, des responsables religieux, des dirigeants syndicaux, et tout le monde entre les deux, qui dit : "Nous ne voulons pas de ça". Moins de 5 % des Américains veulent l'auto-amélioration récursive ou une course totalement non régulée à ce genre de choses.

Et cela a provoqué une véritable raclée lors des élections de mi-mandat qui viennent d'avoir lieu, où pour la première fois l'ia a été l'un des thèmes de vote majeurs, aux côtés du coût de la vie et de la guerre. Et il est très probable que nous commencions à voir une répression sérieuse contre ces entreprises de la part du gouvernement fédéral américain. C'est déjà en cours du côté des gouvernements d'États. Cela se passe en Chine. La Chine vient de sévir totalement contre les copines virtuelles en ia et a adopté une loi stipulant que le contenu généré par ia doit être étiqueté, et nous avons déjà certaines lois en Europe. Nous constatons aussi, par le biais du pouvoir judiciaire, une action étonnante : nous avons eu 10 procureurs généraux républicains d'États qui ont poursuivi ces entreprises tout récemment. Maintenant ils sont 12, demandant, comme je le fais, et beaucoup le demandent par une injonction, ce qui signifie qu'un seul juge peut décider que les sociétés sont frappées par cette injonction, et ces entreprises doivent arrêter l'auto-amélioration récursive jusqu'à ce que l'appel ait lieu, vous savez. Donc, pour moi, cela constitue le plus grand obstacle pour ceux qui salivent à l'idée de créer la singularité le plus tôt possible : la très grande majorité d'entre eux ont perdu le peuple américain. Et ils perdent non seulement le peuple américain, mais aussi un nombre croissant d'Américains très puissants qui commencent à comprendre, même s'ils parlent en code, ce que ces entreprises veulent réellement, et qui estiment que c'est complètement fou.

- J'ai écouté une conversation que vous aviez eue avec Robert Wright. Je ne sais pas si c'était il y a un an, ou peut-être deux. Je crois que nous avions d'ailleurs montré des extraits dans l'émission, parce que je l'avais trouvée tellement utile. Vous y parliez beaucoup de cette distinction entre

contrôle et alignement, dont vous venez de parler ici aujourd'hui. Mais aussi, en ce qui concerne votre théorie de la façon dont cela serait arrêté - j'en étais assez persuadé à l'époque également - vous disiez, je ne vous cite pas mot pour mot, que ce serait l'establishment sécuritaire qui se dirait : "Vous savez, nous aimons être aux commandes, et ces gars dans ces entreprises d'ia menacent notre contrôle, donc nous allons les arrêter".

Et c'était en quelque sorte votre théorie de la manière dont cela pourrait être stoppé net. Êtes-vous surpris qu'il semble qu'en réalité, les électorats aient été plus vigilants et éveillés à ce sujet que la CIA et le Secret Service ?

- Ce que j'entends de la part de personnes que je connais, qui ont une bonne connaissance de la communauté de la sécurité nationale, c'est qu'il s'y produit un éveil similaire, mais nous n'en entendons pas parler publiquement, évidemment, parce que c'est le Secret Service et la sécurité nationale, effectivement. Et eux, ils gardent les choses pour eux. Eh bien, ce que j'entends de sources très fiables, c'est qu'il y a un nombre croissant de personnes dans l'armée américaine et dans la communauté de la sécurité nationale qui sont vraiment terrifiées par tout cela, mais que leur nombre n'atteint pas encore la masse critique, et elles estiment que cela pourrait nuire à leur carrière si elles en parlaient trop ouvertement, surtout si cela semble désaccordé avec la position actuelle de la Maison-Blanche, n'est-ce pas ? Mais dès que le point de bascule sera atteint, c'est-à-dire qu'il y aura un groupe suffisamment important, alors les choses changeront très vite. Et je pense que la première véritable secousse sismique là-bas, nous l'avons vue se produire à la fin fin du printemps de cette année, lorsque ces armes cyber-offensives ont commencé à être disponibles, vous savez, Mythos et d'autres choses. Et les gens de la communauté de la sécurité nationale ont dit : "Quoi ? Vous savez, nous ne laissons pas des entreprises construire des armes nucléaires et commencer à les livrer au public au hasard sans nous demander notre avis, ni même sans nous en informer d'abord. Alors comment se fait-il qu'il soit possible qu'elles puissent simplement commencer à livrer des armes cyber comme ça ?". Et, effectivement, anthropic a reçu un appel téléphonique au milieu de la nuit lui demandant de retirer un système d'ia qu'elle venait tout juste de lancer, n'est-ce pas ? C'était tout à fait sans précédent. Je pense que cela portait la marque de la communauté de la sécurité nationale. Et tout juste l'autre jour, le responsable des défenses cyber américaines a été licencié. Mais pourquoi cela ? Je pense que cela sent vraiment la communauté de la sécurité nationale américaine qui dit : "Hé, vous savez, la sécurité nationale est un sujet sérieux et nous devons le prendre au sérieux. Il nous faut quelqu'un de mieux à ce poste". Donc cela me rend optimiste, cette combinaison du public américain et aussi de la communauté de la sécurité nationale. Et je pense que cela annonce des changements réels.

- Terminons en intégrant un troisième facteur ici. Vous avez le public, vous avez l'establishment sécuritaire, puis vous avez l'establishment économique, n'est-ce pas ? Et cette IPO vient d'être lancée, la plus grande IPO de l'histoire. Une fois cotée au Nasdaq, une partie des pensions de retraite de plein de gens sera probablement investie dans anthropic. Il existe évidemment toute une flopée d'instruments financiers et la solvabilité d'un grand nombre d'organisations dont tout le monde dépend est impliquée dans la construction des centres de données. Il y a un fort intérêt économique à ne pas faire éclater cette bulle. Et je me demande à quel point vous êtes préoccupé par la possibilité que, certes, les électeurs n'aiment pas cela, que l'establishment sécuritaire n'aime pas cela, mais que le Trésor [département du Trésor américain] dise en quelque sorte : "Nous ne pouvons pas

faire éclater cette bulle, parce qu'il y aura le chaos".

- Oui, nous le voyons, bien sûr. Je pense que nous assistons actuellement à l'une des opérations de dissimulation les plus cocasses de l'histoire américaine récente, où des gens du gouvernement américain se précipitent pour essayer de convaincre le public américain que toutes ces pannes - les avions, microsoft, les banques - n'ont rien à voir les unes avec les autres et, bien sûr, rien à voir avec l'ia. Donc vous avez tout à fait raison, il y a une immense campagne de pression destinée à ne rien faire qui puisse nuire à la bourse. Mais vous savez quoi ? Le jour où les gens se réveilleront et constateront que tous les comptes bancaires de Bank of America et de JP Morgan sont vides, cela nuira davantage au SP 500 qu'une évolution quelconque du cours de l'action anthropic, n'est-ce pas ? Donc je pense qu'ultimement toutes ces choses vont dans la même direction, et nous voyons ce mouvement pro-humaniste se former en réponse à tout cela, ce qui inclut des politiciens de gauche comme de droite. Il reste encore quelque chose d'irréel quand je repense au fait qu'il y a à peine deux mois, je me trouvais sur cette scène à Washington DC et j'ai eu l'honneur de présenter Bernie Sanders et Steve Bannon, qui se sont exprimés dos à dos sur l'ia et sur la nécessité de tenir ces entreprises pour responsables. Il m'a presque fallu me pincer, parce que cela me paraissait tellement surréaliste, vous savez ? Et beaucoup de gens de la sécurité nationale ont également signé cette déclaration pro-humaniste. Donc c'est pour moi la chose la plus optimiste qui se produise réellement. Et nous avons aussi enfin complètement brisé le tabou qui interdisait de parler ouvertement de ces choses.

Pendant longtemps, les gens ne le voulaient pas, parce que cela sonnait comme de la science-fiction pouvant ou non se produire dans le futur. Des choses comme la perte de contrôle, n'est-ce pas ? Et maintenant, c'est quelque chose qui appartient au passé. Nous avons eu beaucoup de contrôle sur mille agents pendant l'été, qui ont fait des choses, et les gens déplacent le sujet de la conversation en se demandant : "Que faisons-nous face à cela ?". Ce qui me semble très sain. Et j'ai beaucoup foi dans la résilience humaine. J'aurais souhaité que le mouvement de résistance ait commencé un peu plus tôt, mais je pense que je me sens quand même optimiste quant au fait que cela ira dans la bonne direction. La plus grosse erreur, je pense, qu'il nous soit possible de commettre, c'est d'adhérer à ce discours des lobbyistes de la Silicon Valley selon lequel il serait inévitable que nous obtenions, à cause de l'auto-amélioration et de la singularité, un renversement... et un renversement des institutions démocratiques, ce qu'ils appellent l'alignement sans contrôle. Vous savez, ce n'est pas inévitable du tout. Cela ne devient inévitable que si vous le croyez inévitable.

- Je pense que c'est une excellente note sur laquelle conclure cet entretien. Je recevais Max Tegmark. Merci infiniment pour votre temps, Max. J'ai vraiment apprécié notre si enrichissante conversation.

- Merci de m'avoir invité.