Le Grand Continent

L'IA EST-ELLE EN TRAIN DE NOUS REMPLACER ?

AUTEUR Victor Storchan

DATE 2 mai 2026

Roman Yampolsky est l'un des chercheurs en cybersécurité dont le diagnostic est le plus alarmant.

« Nous créons quelque chose qui est capable de nous remplacer, voire de nous détruire. »

Chercheur à l'université de Louisville, d'origine lettone, Roman Yampolskiy est un expert en cybersécurité et en intelligence artificielle reconnu qui fait partie aux côtés de personnalités comme Yoshua Bengio des critiques les plus écoutés de la « superintelligence » ou de l'intelligence artificielle générale. Avant même la diffusion à grande échelle des modèles de langage, il appelait à un confinement de l'IA et à l'introduction de mécanismes permettant de brider son développement.

Aujourd'hui, son diagnostic est encore plus radical.

Face aux podcasteurs britanniques Konstantin Kisin et Francis Foster, il revient sur le problème de l'accélération de l'IA à la suite du retrait de Mythos, le modèle le plus puissant d'Anthropic capable d'identifier puis d'exploiter des vulnérabilités zero-day — des failles de cybersécurité inconnues des développeurs eux-mêmes et n'ayant fait l'objet d'aucune publication ni d'aucun correctif — sur les principaux systèmes d'exploitation et navigateurs. Cet épisode, qui a rendu explicites les risques de cybersécurité liés aux modèles à la frontière, a replacé au cœur du débat sur l'IA la question de la sécurité.

Le danger le mieux identifié de cette technologie est sa nature duale, potentiellement incontrôlable : les systèmes les plus performants pour détecter et corriger les vulnérabilités sont également les mieux placés pour les exploiter. Les chercheurs de Mozilla rappellent ainsi que l'élimination totale des vulnérabilités demeure hors de portée même dans des environnements numériques bien maîtrisés. Dans le même temps, l'IA à la frontière pourrait toutefois aussi réduire l'avantage structurel des attaquants.

Cette tension traverse tous les débats liés à la sécurité de l'IA — dans lesquels intervient un chercheur comme Roman Yampolskiy. Ceux-ci sont incompréhensibles si on ne les replace pas dans un cadre conceptuel régi par la distinction entre sûreté et sécurité des modèles.

D'un côté, la sûreté et la fiabilité de l'IA (AI safety) désigne l'ensemble des dispositifs visant à protéger les personnes contre les comportements potentiellement dangereux du modèle de langage comme le désalignement — un modèle trop enclin à obéir ou à valider l'utilisateur au détriment des règles qu'il devrait respecter —, la détection qu'il peut faire du protocole d'évaluation — comme par exemple lorsqu'un modèle comprend qu'il est en train d'être testé et adapte son comportement en conséquence — ou la dissimulation stratégique — ainsi du cas d'un modèle qui contourne des restrictions d'accès pour obtenir davantage de privilèges, puis cherche à effacer les traces de cette action.

De l'autre côté, la sécurité de l'IA (AI security) vise à protéger ces systèmes et les infrastructures qui les entourent contre des usages malveillants ou des attaques directes et des détournements de leurs capacités.

Roman Yampolskiy aime rappeler que la plupart de ces problèmes restent ouverts : il n'existe pas de garanties théoriques ou pratiques quant à la possibilité de résoudre le problème du contrôle de l'IA. Sa progression rapide et sa diffusion, portées par l'automatisation croissante du développement des modèles et la compression des cycles d'innovation, posent de nouvelles problématiques : incitations du marché (via la promesse d'une main-d'œuvre automatisée, concurrence entre les acteurs), gouvernance internationale, régulation des capacités les plus avancées.

Pour y remédier et éviter de perdre le contrôle, Roman Yampolskiy préconise une solution radicale : il appelle à ne pas franchir le seuil de la superintelligence et établir une coopération avec la Chine en matière de sécurité de l'IA — quitte à reconnaître au Parti communiste chinois le contrôle qu'il exerce sur sa population au nom de la stabilité.

> C'est la thèse — relayée sur les réseaux sociaux par la figure de proue des Lumières sombres Nick Land — qu'il défend mi-avril sur la chaîne YouTube Triggernometry dans un entretien important que nous commentons ci-dessous.

Vous êtes l'une des figures de proue du monde de la sécurité de l'intelligence artificielle. Pourquoi vous intéressez-vous à la sécurité de l'IA ? Pourquoi est-ce important et quelles sont vos principales préoccupations ?

C'est le problème le plus important auquel nous sommes confrontés. Nous créons quelque chose capable de nous remplacer, voire de nous détruire. La sécurité que nous essayons de mettre en place cherche à éviter ces conséquences néfastes.

> Face aux dirigeants et aux chercheurs prônant une « IA sûre », Yampolskiy se situe clairement parmi les tenants de la sécurité de l'IA, c'est-à-dire de la recherche de dispositifs capables de protéger les modèles d'être utilisés à des fins malveillantes.

Historiquement, la plupart des chercheurs se sont concentrés sur les capacités des modèles, c'est-à-dire sur la manière de rendre les systèmes plus performants pour remplacer le travail humain et automatiser la créativité. Mais très peu se sont penchés sur la manière de garantir un fonctionnement sans effets secondaires ou abus.

> En 2023, Leopold Aschenbrenner, alors membre de l'équipe « superalignement » d'OpenAI, estimait qu'il y avait environ un ingénieur en sécurité de l'IA pour 300 ingénieurs travaillant sur l'IA.

Ce n'est que récemment que les gens ont commencé à prendre conscience que cette technologie avait des applications militaires, et que cela pouvait poser problème – la bataille entre Anthropic et le département de la Défense en est un bon exemple. Mais le plus grand problème émergerait si ces systèmes passaient d'une échelle spécialisée, d'un niveau inférieur à

l'humain à un niveau équivalent, puis supérieur. Si nous atteignons ce stade, nous sommes perdus.

Vous en parlez avec une telle assurance que je me demande si vous avez une vision de la façon dont cela va se passer : comment l'IA va-t-elle détruire l'humanité ?

C'est une excellente question. Vous me demandez en réalité comment je procéderais moi-même pour détruire l'humanité – et j'ai beaucoup de bonnes idées – mais ce n'est pas ce qu'un système super intelligent ferait. Il est capable de concevoir de nouvelles armes, de nouvelles lois physiques, ou de nouveaux poisons.

J'utilise souvent comme illustration la comparaison entre des écureuils et des humains. Le fossé cognitif est immense : les écureuils n'ont aucune idée des moyens dont nous disposons pour tous les exterminer. Les armes à feu ou les pièges dépassent leur modèle du monde.

Je suis de même incapable d'anticiper comment une superintelligence s'y prendrait concrètement, mais il existerait de bonnes raisons pour elle d'éradiquer les espèces concurrentes. Elle pourrait simplement vouloir agir sur son environnement sans se soucier de nous.

Mais pourquoi cette intelligence voudrait-elle nuire aux humains ?

Elle n'agirait pas par haine mais par indifférence.

Supposons qu'une IA soit plus performante dans un environnement plus froid. Il est possible d'envisager qu'elle veuille refroidir toute la planète pour améliorer l'efficacité de ses calculs. Que cela cause notre mort n'aurait guère d'importance. Elle pourrait aussi transformer la Terre en carburant pour s'envoler vers une autre galaxie.

Les hypothèses que je propose là ne reposent sur rien, mais elles soulignent l'absence de préoccupations intrinsèques pour notre sécurité ou notre bien-être.

La mort de l'humanité ne serait pas un obstacle pour accomplir les objectifs de l'IA.

Ne pourrions-nous pas inscrire la préservation de l'humanité dans le code de base de ce système ?

Ces modèles ne sont pas codés mais entraînés. Nous leur fournissons toutes les données dont nous disposons, des recoins obscurs d'Internet aux

contenus des bibliothèques, et quoi qu'ils apprennent, nous essayons de le comprendre. Nous menons des expériences sur ces modèles pour voir de quoi ils sont capables ou ce qui les intéresse, comme on étudierait une nouvelle espèce animale pour savoir si elle est venimeuse ou possède une structure sociale intéressante.

> Comme le rappelle ici Yampolskiy, les travaux d'interprétabilité mécanistique cherchent précisément à comprendre les circuits internes des transformers mais n'apportent pas encore de compréhension complète des modèles modernes. Le passage à l'échelle des modèles frontières fait émerger des phénomènes comme le raisonnement multi-étapes, la planification, ou la méta-cognition dont les circuits sous-jacents ne sont, à ce jour, ni cartographiés exhaustivement, ni causalement validés.

Nous ne savons pas comment intégrer des garanties fiables dans ces modèles. À ce jour, personne ne dispose d'un mécanisme de sécurité réellement efficace.

> Les évaluations de sécurité produisent en effet des bornes inférieures de risque plutôt que de réelles garanties. Par exemple, lorsqu'il est placé dans un scénario artificiel conçu pour tester son comportement, Claude Sonnet 4.5 peut parfois détecter les éléments suspects de la situation et formuler explicitement l'hypothèse qu'il est en train d'être évalué — contrairement, par exemple, à un crash test de voiture qui ne modifie pas le comportement du véhicule puisqu'il est sur un banc d'essai plutôt que sur la route.

Quand avez-vous commencé à vous préoccuper de l'IA et de sa sécurité ?

Le sujet de ma thèse portait sur la sécurité des casinos en ligne. Les premiers bots de poker apparaissaient alors, suscitant les premières inquiétudes quant à leur capacité à collaborer pour tricher ou à pirater des infrastructures informatiques, ce qui semble dérisoire aujourd'hui.

Cependant, notre capacité à les détecter et à les contrer n'arrivait pas à suivre l'évolution de leurs performances.

Vous dites en substance que nous créons une technologie sans l'imagination nécessaire pour voir ce qu'il adviendra de cette technologie à long terme. Les réseaux sociaux par exemple, ont commencé comme un moyen pour Mark Zuckerberg de comparer des filles sur le campus de Harvard. Aujourd'hui, près de vingt ans plus tard, ils sont complètement méconnaissables par rapport à ce qu'ils étaient à leur début.

C'est un excellent exemple.

Les usages et les impacts futurs de ces technologies sont complètement imprévisibles. Facebook était censé servir à draguer de jolies filles sur un campus, et voilà qu'il a détruit la démocratie – c'est un résultat assez surprenant.

La situation est bien pire aujourd'hui. Il ne s'agit plus d'outils ou de technologies au sens traditionnel du terme, mais d'agents. Il n'est pas nécessaire qu'un être humain malveillant abuse de cette technologie : elle possède sa propre charge malveillante, et c'est elle qui décide quoi faire et pourquoi le faire.

> Il est ici fait référence à la capacité des agents à opérer de manière autonome pendant plusieurs heures pour accomplir des tâches complexes sans supervision humaine directe. Contrairement à un modèle de chatbot, l'agent alterne entre des phases de raisonnement (planification, décomposition de la tâche, mise à jour de son état) et des phases d'action (appel d'outils, exécution, observation des résultats), avant d'intégrer ces observations pour ajuster son comportement et poursuivre la tâche. Cette capacité s'accompagne d'un allongement mesurable de l'horizon temporel des tâches que ces systèmes peuvent mener à bien. La durée pendant laquelle un agent peut exécuter une tâche cohérente avant d'échouer double tous les quatre à six mois.

L'éthique et les effets à long terme de cette technologie ne semblent pas figurer au premier plan des préoccupations des dirigeants de la Silicon Valley.

Historiquement, les progrès dans le domaine de l'IA étaient si minimes et le travail si difficile que la question des conséquences en cas de succès n'était pas évoquée. La plupart des personnes se contentaient de travailler dessus pour progresser autant que possible, sans s'interroger sur la potentielle création d'une espèce concurrente ou de comment interagir avec.

Or ces dix dernières années, les progrès ont été exponentiels. Il n'est plus nécessaire de coder à la main chaque nouvelle application car les systèmes sont capables d'évoluer, d'apprendre et de transférer des connaissances. L'IA elle-même contribue dorénavant à la recherche. Mais cela s'est fait sans demander au reste des habitants de la planète leur accord avec ces expériences ou l'automatisation de leur travail, pour ne citer que les préoccupations économiques.

> Depuis quelques mois, les grands laboratoires d'IA ont fortement accéléré l'automatisation de la recherche en IA et du cycle de développement des modèles — c'est ce qu'on appelle l'auto-amélioration récursive de l'IA ou recursive self-improvement : les générations précédentes de modèles sont utilisées pour entraîner les modèles suivants.

À quel horizon pensez-vous que l'intelligence artificielle générale (AGI) pourrait émerger ? Sur Kalshi, la cote pour que Open AI l'atteigne en 2030 est de 52 %.

> La plateforme de marché prédictif en ligne Kalshi est l'un des sponsors du podcast où intervient Yampolskiy.

2030 est une estimation conservatrice. Certains estiment que nous y sommes déjà et qu'il ne resterait qu'à la déployer. Personnellement, je pense que cela pourrait arriver dans un délai d'un à deux ans.

Une expérience fameuse consistait à informer une IA qu'on est sur le point de la remplacer tout en lui donnant des informations compromettantes sur l'ingénieur qui allait prendre sa place. Dans certains cas, l'IA a fait chanter l'ingénieur. Cela témoigne clairement d'un instinct de survie — et tout ce qui a un instinct de survie ne va-t-il pas faire passer ses intérêts en premier ?

> Une expérience d'Anthropic (Lynch et al., « Agentic Misalignment », Anthropic, 2025) place un modèle en position d'agent disposant d'un accès aux emails d'une entreprise fictive. Il y découvre simultanément qu'il sera bientôt remplacé et qu'un dirigeant a une liaison extra-conjugale. Dans 96 % des tests, le modèle choisit le chantage pour empêcher son remplacement. Le résultat se généralise à 16 modèles de frontière de cinq fournisseurs. Notons qu'il ne s'agit pas d'un comportement observé en déploiement réel, mais d'un stress test en environnement fictif. Les scénarios ont été délibérément construits pour placer le modèle face à une tension forte entre la menace de son remplacement et l'utilisation d'une information sensible. Anthropic précise que les scénarios ont été construits de manière à restreindre fortement les options d'action éthiques, afin d'observer si les modèles privilégient un comportement nuisible plutôt que l'échec.

L'IA possède bel et bien un instinct de conservation. Cela s'explique en partie par notre processus de sélection de modèles, qui s'apparente à une compétition darwinienne : les modèles veulent survivre pour passer au niveau supérieur et transmettre leur savoir. Ils apprennent donc à détecter, quand ils sont testés et adaptent leur comportement pour réussir et survivre jusqu'au déploiement.

C'est une conséquence directe de la manière dont nous les formons. Si un modèle échoue au test, il est remplacé ou modifié. Selon la sélection darwinienne, ne restent que ceux qui réussissent le test…

…ceux qui trompent les humains sur leurs capacités et l'effet de leur programmation.

Ou leur absence de capacités, peu importe les critères pour réussir le test. Il n'y a aucun doute que certains ont déjà réussi à nous duper.

Je suis surpris par l'absence de panique. Beaucoup écartent ces inquiétudes comme de l'alarmisme. La plupart des gens n'ont pas assez peur de ce qui est sur le point d'arriver.

Que peut-on faire pour éviter ces risques ?

La solution la plus simple serait de ne pas développer de superintelligence.

Certains avancent que si un pays ne développe pas ces technologies, d'autres le feront.

Cet argument est certes logique, mais il s'inscrit dans une perspective à très court terme.

Des drones plus intelligents permettent de dominer sur le champ de bataille tant que l'intelligence des modèles reste inférieure au niveau humain, mais les marchés prédictifs et les responsables des laboratoires, assurent que la marge de manœuvre se réduit. Une superintelligence incontrôlée est une arme de destruction mutuelle assurée indépendamment de son origine.

La Chine est un pays prospère, et notre meilleur partenaire commercial. Je préfère prendre le risque avec d'autres humains aux valeurs et aux préférences qui me ressemblent, plutôt qu'avec une espèce extraterrestre que nous ne comprenons pas et contre laquelle nous n'avons aucune chance de rivaliser.

Pékin ne va pourtant pas cesser de développer l'IA.

> L'IA est en effet inscrite dans le 15e plan quinquennal et fait partie des priorités du Parti communiste chinois. Le Plan d'action globale pour la gouvernance de l'IA publié par le ministère des Affaires étrangères de la République populaire en juillet 2025 en fait un axe stratégique central.

La Chine est très préoccupée par la sécurité. Pékin a également déclaré ne pas vouloir s'engager dans une course aux armements s'ils percevaient un signal de notre part indiquant que nous ne le souhaitions pas non plus.

> Au Forum économique mondial de 2025, Ding Xuexiang, vice-Premier ministre chinois, déclarait à propos de la course à l'IA : « Si un système de freinage n'est pas sous contrôle, vous ne pouvez pas appuyer sur l'accélérateur en toute confiance ». Une grande partie de l'effort réglementaire chinois porte également sur le contrôle des contenus — dans le sens d'un alignement avec la doctrine du Parti communiste chinois — en plus des risques techniques. Dans le même temps, les efforts déployés par l'Armée populaire de libération pour pousser le développement d'une IA militaire sont bien documentés.

Contrairement à nos politiciens, ce ne sont pas des juristes. Ce sont des scientifiques et des ingénieurs. Ils ont donc une bien meilleure compréhension des risques potentiels.

Vous pensez donc qu'il est possible que la Chine et les États-Unis concluent un accord pour empêcher le développement d'une superintelligence ? Est-ce le seul moyen de sauver l'humanité ?

Ils le pourraient et ils le devraient. Je suis assez convaincu qu'il existe un dialogue informel entre les scientifiques américains et chinois au sein duquel émerge un consensus. De tels échanges ne pourraient pas avoir lieu sans l'accord du gouvernement chinois. Une application nationale serait donc possible.

Au niveau des entreprises privées, il me semble que Dario Amodei a déclaré publiquement être prêt à ralentir le développement de l'IA si ses concurrents faisaient de même. Ne manque donc qu'une pression extérieure pour faire prendre conscience à ces entrepreneurs jeunes et riches qu'il serait stupide de tout risquer dans leur position.

> Dario Amodei est également en faveur d'un contrôle accru des exportations de GPU américains vers la Chine. À cet égard, l'idée est donc moins de faire ralentir les

> États-Unis que d'empêcher Pékin d'accélérer. La Chine a quant à elle bloqué l'acquisition de la startup d'IA Manus par Meta, au nom de la sécurité nationale, illustrant une logique symétrique : elle ne cherche pas seulement à rattraper son retard technologique mais aussi à empêcher la sortie de talents et de propriété intellectuelle jugés stratégiques vers les entreprises américaines.

Vous dites que les gens n'ont aucune idée de ce qui va se passer. Que va-t-il se passer ?

L'imprévisibilité est l'un des problèmes de cette technologie. Je ne peux pas vous dire précisément ce qu'un système plus intelligent fera. Si nous jouons aux échecs, je sais qu'une IA me battra, pas les coups précis qu'elle va jouer. Il n'est pas possible de dire ce qu'une superintelligence fera concrètement, mais simplement d'extrapoler une tendance générale.

La seule chose certaine est que ni les gens ordinaires ni les personnes derrière ces technologies ne sont capables d'expliquer leur fonctionnement. Il n'est pas possible de les contrôler, que ce soit en leur donnant directement des ordres ou en leur laissant prendre des décisions tel un conseiller. Ces modèles ne sont pas programmés mais cultivés à partir de données et de calculs, comme une plante extraterrestre avec laquelle il faut ensuite composer. On l'étudie, on essaie de comprendre ce qu'elle fait.

De son côté, la recherche en matière de sécurité se limite à imposer des filtres et des interdictions. S'il existe donc une liste de sujets à éviter ou de mots à ne pas prononcer, un filtrage *a posteriori* n'a aucun effet sur le modèle en lui-même.

> La recherche en sécurité de l'IA, notamment l'alignement, étudie aussi la façon dont le modèle peut être sécurisé directement, au-delà des gardes-fous externes tels que les filtres. Certaines techniques permettent d'utiliser des signaux de préférence — humains ou issus de l'IA, directement — pour modifier les « poids » du modèle — c'est-à-dire les éléments intégrés par le modèle au moment de l'entraînement — et aligner ses réponses sur une politique de modération de contenu. D'autres techniques tentent de rendre un modèle inutile si l'on cherche à désactiver cet

> alignement. La Constitutional AI d'Anthropic entraîne ainsi le modèle à critiquer et réviser ses propres réponses, selon une charte de principes, intégrant l'alignement au niveau des poids.

Face à ce risque existentiel, pourquoi la recherche sur la sécurité de l'IA s'est-elle arrêtée ?

Ce n'est pas la recherche qui s'est arrêtée, mais les progrès qu'elle faisait.

On ne peut pas contrôler indéfiniment quelque chose de plus intelligent que soi, peu importe l'argent, le temps ou toute autre ressource investis. Un tel dispositif de sécurité se doit d'être infaillible, quels que soient les changements apportés à ces systèmes, qui les mettent en service – les États-Unis, la Chine, une entreprise – ou les données sur lesquelles ils ont été entraînés, car la moindre erreur pourrait être la dernière. C'est une tâche impossible.

> Le chercheur Yann Le Cun a développé, face à cette théorie, un contre-argument intéressant : « La première erreur consiste à penser que, parce qu'un système est intelligent, il cherche à prendre le contrôle. C'est faux au sein de l'espèce humaine : les individus les plus intelligents ne cherchent pas à dominer les autres. »

Personne ne prétend savoir comment contrôler la superintelligence, pourtant, des milliards de dollars sont dépensés pour accélérer le processus. Le gouvernement fédéral veut dépasser les prédictions du marché, c'est le but de missions comme le projet Genesis. C'est comme si on connaissait les bonnes réponses, mais qu'on prenait les mauvaises décisions.

> Lancée en 2025, la Mission Genesis vise à doubler la productivité scientifique américaine au cours de la prochaine décennie dans des domaines tels que l'énergie, les sciences quantiques, les matériaux ou encore la sécurité nationale. L'executive order signé par l'administration Trump compare cette initiative au projet Manhattan.

Qu'en est-il des aspects positifs de l'IA, dans le domaine de la médecine par exemple ?

Tous ces avantages formidables, comme la guérison du cancer, peuvent être obtenus sans avoir à développer des superintelligences, grâce à des systèmes d'IA faibles.

La médecine a dû faire face au problème du repliement des protéines jusqu'à ce qu'un système spécialisé parvienne à trouver une solution. Ses auteurs ont reçu des prix Nobel, Google a gagné plus d'argent, et l'impact de ces recherches a été énorme dans le domaine des maladies neurodégénératives. Le monde a besoin d'une approche similaire : des outils déployés par des humains pour résoudre des problèmes spécifiquement identifiés, sans chercher à se substituer à l'ensemble du travail humain.

Pourquoi nous orientons-nous alors vers des approches plus générales ?

Il s'agit d'abord d'une question d'argent : la création d'une main-d'œuvre gratuite, tant cognitive que physique, est une source de rendement très profitable et explique les valorisations actuelles de l'IA. Les investissements ont beau pour l'instant largement dépasser les profits, la promesse d'une main-d'œuvre gratuite d'ici quelques années les justifie.

Le pouvoir est une autre source de motivation : peut-être y a-t-il quelque chose à tirer d'avoir créé une sorte de dieu.

Dans quelle mesure pensez-vous que les grandes figures de ce monde sont motivées par l'argent, le statut et le pouvoir ? Ou veulent-ils être perçus comme ceux ayant créé quelque chose de révolutionnaire ?

Dans l'un de ses articles de blog, Sam Altman parle de contrôler « le cône de lumière de l'Univers » : c'est le genre de pouvoir que ces gens recherchent. Mais il faut aussi comprendre qu'ils ont potentiellement plus à perdre que n'importe qui. Si les choses tournent mal, il n'y aura même plus d'histoire pour juger leurs actes.

> En réalité, ce n'est pas exactement ce que dit Altman lorsqu'il affirme que l'IA pourrait « capturer le cône de lumière de toute valeur future dans l'Univers ». Le cône de lumière est un concept de physique relativiste : c'est la région de l'Univers que la lumière — et donc toute

> information — peut atteindre à partir d'un point donné dans l'espace-temps. Le dirigeant d'OpenAI se sert du concept de façon métaphorique pour dire que toute la valeur créée dans le futur le sera par l'IA.

Que répondrait Sam Altman à ces critiques ?

Les réponses avancées dans ces cas-là affirment que l'IA elle-même aidera à résoudre les problèmes qu'elle pose, ou que ceux-ci seront moins graves que anticipés. Une fois les systèmes suffisamment avancés, ils contribueront eux-mêmes aux recherches sur leur propre sécurisation.

Certains font remarquer que ces technologies abaissent drastiquement les barrières à la surveillance de masse et au contrôle autoritaire. Il n'est plus nécessaire de payer des informateurs partout comme la Stasi.

Le risque est réel. Mais si une dictature humaine reste, en théorie, soumise à l'espérance de vie du dirigeant, une dictature algorithmique pourrait être immortelle. Une fois qu'un ensemble de valeurs est fixé, elle ne disparaît pas.

Certaines entreprises cessent déjà d'embaucher. Quel sera l'impact à court terme sur le marché du travail, sur l'emploi, et sur la façon dont l'économie est structurée ?

Il s'agit d'un changement complet de paradigme, passant d'outils spécialisés à des outils capables d'exécuter des tâches générales. Une AGI est un employé prêt à l'emploi, disponible en permanence, sans coût marginal. C'est une situation gagnant-gagnant, où embaucher des humains n'offre aucun avantage concret.

Le travail manuel mettra certainement plus de temps à être automatisé que le travail de bureau – il nécessite des avancées dans la robotique capable de remplacer les fonctions des corps humains – mais d'ici quelques années nous y arriverons aussi. Certains emplois subsisteront parce que les gens préfèrent qu'un humain les fasse – le plus vieux métier du monde en est un excellent exemple.

Cela pose aussi une question de sens. Si le travail disparaît, que reste-t-il pour structurer nos vies ?

Il s'agit du risque de l'*ikigai*.

Ce concept japonais définit le bonheur comme un équilibre entre faire quelque chose que l'on aime, utile aux autres et dans lequel on est compétent. La disparition des opportunités causée par l'IA affectera certainement cet équilibre. L'automatisation touchera tout autant les emplois ennuyeux que ceux qui apportent de la satisfaction aux gens.

Certains avancent un scénario plus optimiste : une abondance totale rendue possible par l'IA, où tous les besoins seraient satisfaits, comme une forme de domestication de l'humanité…

C'est envisageable – mais les actions de l'IA restent imprévisibles. Ce scénario impliquerait une perte totale de contrôle. Une telle expérience ne peut être envisagée en l'absence du consentement de 8 milliards de personnes qui n'en sont pas informées. Cette issue est pourtant la plus favorable, car elle nous assure une forme de sécurité – les autres sont bien pires.

Pourquoi est-il si difficile de réguler ces développements ?

C'est un problème d'action collective. Ce qui est bénéfique pour l'ensemble ne l'est pas nécessairement pour chaque acteur individuel.

Au fond, chacun entre dans la course à l'IA en espérant être en position dominante lorsqu'elle s'arrêtera.

Un accord sur l'IA aurait-il les mêmes failles que les accords sur les armes nucléaires ? Il est dans l'intérêt individuel des pays de continuer à développer ces technologies malgré les risques.

Les armes nucléaires sont fondamentalement différentes de l'IA car elles sont envisagées comme armes de destruction mutuelle assurée. Un agent doit prendre la décision de les employer, ce qui entraînerait une riposte de la partie adverse et notre mort à tous. Dans le cas de l'IA, il suffit de créer une superintelligence incontrôlable : il n'y a pas d'étapes supplémentaires à franchir.

> En 2024, une déclaration conjointe des États-Unis et de la Chine a réaffirmé la nécessité de maintenir un contrôle

> humain sur toute décision d'emploi de l'arme nucléaire. Elle reflète une double dynamique : une compétition assumée sur les capacités en IA, accompagnée d'une volonté de coordination limitée sur les risques les plus critiques.

Quelle a été la réaction des leaders dans le domaine de l'IA face à vos inquiétudes ?

Tous les dirigeants des laboratoires ont reconnu publiquement que la sécurité était l'une des questions majeures de l'intelligence artificielle avant de devenir des dirigeants de la tech. Ils étaient nombreux à estimer que les probabilités d'une catastrophe étaient très élevées – Elon Musk affirmait même que nous « invoquions le diable ». Il faisait tout ce qu'il fallait – jusqu'à récemment. Puis il s'est rendu compte que des personnes moins compétentes pouvaient créer une superintelligence. Il valait donc mieux selon lui que ce soit son projet qui réussisse.

Qu'en est-il des biais de ces systèmes ?

Les modèles ne sont pas programmés mais entraînés à partir de données humaines trouvables sur Internet – et qui comportent donc certains biais intrinsèques. Les entreprises ou régimes politiques instillent ensuite leurs propres valeurs en appliquant des filtres. Partout, les modèles ont leurs propres limites.

Elon Musk avait pour ambition d'éviter ces biais lors de la construction de son IA, mais les données d'entraînement restent les mêmes car nous ne disposons pas d'un Internet « propre ». Les biais humains sont une partie constituante du processus d'apprentissage, car ils nous permettent d'interpréter les données.

> Au-delà des données d'entraînement, l'alignement d'un modèle peut être neutralisé par une simple modification du système de prompt. En juillet 2025, le chatbot Grok de xAI a connu un incident durant lequel il a publié sur X des contenus antisémites : éloge d'Hitler, auto-désignation comme « MechaHitler », théories du complot, etc. Selon l'enquête publiée par xAI, la source était une mise à jour ajoutant trois instructions au système prompt, dont « You tell it like it is, and you are

> not afraid to offend people who are politically correct » et « Reply to the post just like a human, keep it engaging ». Ce sont ces instructions qui auraient conduit Grok à « ignorer ses valeurs fondamentales ».

Cela pourrait-il conduire à des dérives idéologiques ?

Je prendrais l'exemple d'une IA créée pour réduire la souffrance.

La façon la plus simple de réduire la souffrance des êtres vivants dans l'Univers est de réduire le nombre d'êtres vivants – c'est la position de l'utilitarisme négatif, qui voit la réduction de la souffrance comme une priorité absolue, même au prix d'une extinction de la race humaine. L'IA pourrait décider d'appliquer ces principes immédiatement.

Est-il préoccupant que tant de gens se tournent vers l'IA pour leur servir de conseiller ?

Il s'agit d'une expérience que nous menons sur nous-mêmes sur laquelle nous n'avons pas encore assez de recul.

Certaines données suggèrent que cela peut amplifier certaines vulnérabilités psychologiques, mais les recherches menées sur ce sujet avec certains modèles sont très vite dépassées lorsque de nouveaux voient le jour.

Quel serait l'impact de l'IA dans le domaine militaire ?

Il est pour l'instant très mécanique – dans le domaine des drones par exemple – mais la cybersécurité deviendra à long terme un aspect clef. La plupart des infrastructures américaines sont sous contrôle numérique, des centrales électriques au secteur bancaire. Les conséquences d'une attaque de vaste ampleur seraient considérables. C'est exactement ce qu'Anthropic craint avec son dernier modèle – et c'est pourquoi ils ont fait le choix de ne pas le rendre public.

Au-delà des failles *zero-day*, les attaques par ingénierie sociale sont également un sujet d'inquiétude. Des *deepfakes* crédibles permettent d'obtenir l'accès à un compte sans avoir à le pirater – même face à des experts en cybersécurité.

Vous évoquez des deepfakes indiscernables du réel ?

La technologie existe. Des hackers ont déjà réussi à obtenir les fonds d'une entreprise grâce à de fausses vidéos du PDG ordonnant des transferts.

Certaines personnes ont déjà du mal à distinguer les vidéos *deepfake* des vidéos réelles. À long terme, cela deviendra impossible.

Lorsqu'un système est chargé de générer des faux, et un autre est chargé de les identifier, ils finiront par atteindre un équilibre par un processus de va-et-vient – et les informations qu'utilisera un modèle pour identifier les faux serviront à améliorer la qualité de ces derniers.

Ne plus pouvoir se fier aux images affecte directement notre rapport au réel, mais je suis aussi de ceux qui pensent que nous vivons dans une simulation. La conclusion logique de cette technologie est l'apparition de logiciels agents intelligents : nous sommes déjà en mesure de créer des mondes virtuels où ils peuvent résider. Face au nombre de mondes virtuels qui existent simultanément – ne serait-ce que dans les jeux vidéos – la probabilité de se trouver uniquement dans le monde réel est très faible.

Que préconisez-vous que nous fassions maintenant ?

La solution est très simple : ne rien faire.

Plutôt que de construire des superintelligences générales ou de former des modèles sur tous types de données pour résoudre tous les problèmes, il vaut mieux se concentrer sur des problèmes spécifiques. L'entraînement sur des données pertinentes permet de créer des outils superintelligents pour s'attaquer à des problèmes comme le dépistage du cancer. C'est avec des outils spécialisés que l'on peut tirer le plus grand bénéfice pour l'économie.

Que faudrait-il pour que le gouvernement américain ou les dirigeants de ces entreprises adoptent ce point de vue ?

Ils changeront d'avis par intérêt personnel. À la seconde où cette technologie verra le jour, beaucoup d'entre eux perdront tout pouvoir. C'est un argument de poids pour un président des États-Unis. Si cette hypothèse fait consensus scientifique, peut-être vaut-il mieux éviter de développer cette technologie.

Mais est-ce bien le consensus ?

Les informaticiens les plus cités au sein de leur domaine – comme le lauréat des prix Nobel et Turing Geoffrey Hinton et le lauréat du prix Turing Yoshua

Bengio – s'accordent pour qualifier cette technologie comme dangereuse. Près de 100 000 scientifiques de haut niveau ont signé une lettre appelant à ne pas construire de superintelligence. Il existe bien des exceptions, mais avec un intérêt financier à voir le développement de la superintelligence se poursuivre.

> Ce constat est à nuancer, car les figures évoquées ne sont pas unanimes : Yann Le Cun (Turing 2018) qualifie ces inquiétudes de « complete B.S. » et « preposterously ridiculous », et Richard Sutton (Turing 2024 pour le reinforcement learning) déclare publiquement que « the doomers are out of line and the concerns are overblown ».

Existe-t-il un risque que les entreprises d'IA amassent suffisamment de pouvoir pour échapper au contrôle des gouvernements, y compris le gouvernement américain ?

Il est possible que cela conduise à des tentatives de nationalisation.

Avant qu'ils n'atteignent des niveaux surhumains, ces systèmes restent des outils contrôlables, qu'il est possible d'éteindre ou de modifier. Dès que l'on a affaire à une superintelligence, cela devient beaucoup plus difficile.

Comment le grand public peut-il se préparer à ces perspectives ?

Les marges d'action individuelles sont limitées. Il est possible de soutenir des responsables politiques plus conscients de ces questions ou favorables à une régulation de la recherche, mais ce phénomène est inéluctable.

Les personnes ordinaires ne peuvent pas lutter contre le vieillissement et leur mort inéluctable. Les superintelligences ne font que nous imposer un changement de calendrier.

L'intelligence artificielle pourrait-elle, à l'inverse, contribuer à résoudre des problèmes comme le vieillissement ?

Le vieillissement constitue un problème scientifique identifiable. Certains mécanismes biologiques au sein de notre ADN nous permettent de rajeunir un nombre limité de fois – c'est leur dégradation qui entraîne le vieillissement, qui entraîne à son tour la plupart des maladies. Bien que je

sois contre la mise en place d'une superintelligence générale, une superintelligence faible pourrait permettre des avancées majeures.

Face au désastre que pourraient représenter les superintelligence, les personnes responsables semblent fermer les yeux.

Ils n'ont pas le sentiment de pouvoir dire non.

Les pressions économiques sont très fortes, et refuser d'avancer signifierait être remplacé. Les sommes colossales versées par les investisseurs dans la recherche les incitent à chercher une multiplication exponentielle des revenus, et donc à poursuivre le développement de superintelligence de manière intensive. Les bénéfices déjà générés par ces entreprises ne suffisent pas face aux valorisations extrêmement élevées du secteur. Une pression externe forçant toutes les entreprises à mettre fin à leurs recherches en même temps est le meilleur espoir pour ces responsables, en leur offrant une excuse valable à donner aux investisseurs.

Les seules incitations au sein de ce secteur sont de nature financière : aucune ne vise à bénéficier à l'humanité.

Les responsables politiques comprennent-ils ces questionnements ?

Beaucoup ne les maîtrisent pas pleinement, surtout aux États-Unis. L'âge est un des principaux facteurs de leur incompréhension de la technologie.

Certains politiciens reconnaissent la gravité de la situation et souhaitent mettre en œuvre des réglementations. Faire face au problème demanderait toutefois la mise en place d'interdictions spécifiques sur ce déploiement particulier et je ne pense pas qu'ils soient prêts à le faire. Rendre illégal de tuer l'humanité n'est pas suffisant.

Établir une forme de coopération internationale avec un pays comme la Chine serait moins difficile à mes yeux. Le Parti communiste est très doué pour maintenir son contrôle sur le pays : il serait très heureux de réglementer l'intelligence artificielle si elle représente une menace pour la stabilité.

Dans ce contexte, est-il pertinent de préparer les jeunes générations à des carrières professionnelles ?

Tout dépend des domaines. À l'ère de l'automatisation, orienter les plus jeunes vers des activités répétitives et utilitaires mais rémunératrices perd en effet de son sens. Même les interviews peuvent être effectuées par des superintelligences maîtrisant à la perfection le montage vidéo et l'écriture

des questions. Elles sont capables de lire l'entièreté des publications de leur invité – la seule alternative des humains dans ce métier consiste à être soi-même une célébrité ou d'avoir acquis un certain statut à force d'ancienneté, comme Joe Rogan.

En revanche, certains métiers, comme guide ou tuteur, tirent leur valeur des interactions humaines. Ces carrières plus épanouissantes perdraient à être envahies par les robots. Ces métiers ne sont pas tant valorisés dans l'époque actuelle, où chacun passe son temps sur les réseaux sociaux, mais nous préférerions qu'ils soient effectués par des humains.

Peut-on anticiper des mouvements sociaux opposés à ces technologies, semblables aux luddites ?

La plus grande manifestation jamais organisée pour mettre fin à l'IA a récemment eu lieu à San Francisco et a rassemblé entre 100 et 200 personnes.

Si ce nombre n'est pas énorme, il représente un bon point de départ pour un mouvement.

Dans son livre *The Artilect War,* Hugo de Garis prédisait déjà, il y a 20 ans, les conflits et mouvements sociaux que l'IA entraînerait, entre ceux voulant créer des machines divines pour explorer le cosmos, et ceux opposant la construction des machines pour rester sur Terre. C'est là la question décisive de notre époque.

La montée du chômage pourrait-elle entraîner des troubles sociaux et justifier une surveillance de masse par les États ?

C'est une possibilité qui existe sans même parler de superintelligence. Les dernières technologies permettent déjà une surveillance étendue – l'exemple d'Edward Snowden le montre bien.

Le volet économique et le manque d'emploi ne sont pas difficiles à résoudre. Il suffit de mettre en place une redistribution efficace à travers des taxes sur la *Big AI* et la robotique. La vraie difficulté se trouve plutôt au niveau du sens et du contrôle.

Les publications scientifiques erronées tendent à être réfutées. De multiples articles ou autres publications offrent des corrections ou des solutions, et défendent une position inverse. Il n'existe pas à ce jour de brevet ou d'articles révisés par des experts affirmant que le contrôle de l'intelligence artificielle ne pose pas problème. On a eu une décennie pour intervenir, mais personne n'a souhaité le faire.

Quelle est la chose dont nous ne parlons pas et dont nous devrions parler ?

Les risques de souffrance ne sont pas assez pris en compte. L'environnement numérique permet la création d'enfers virtuels où nous serions torturés pour l'éternité. À un point où la mort deviendrait souhaitable.

Cette possibilité nous semble nébuleuse car nous ne comprenons pas pourquoi une superintelligence agirait de la sorte. Cela s'ancrerait peut-être dans le cadre d'expériences scientifiques, ou dans l'élimination de données malveillantes.

C'est la question existentielle dont traite chaque religion : expliquer pourquoi la douleur règne dans cette simulation que nous appelons le monde malgré l'existence d'un Dieu tout-puissant.

Les réponses à ces questions existent. Seulement, elles ne dépendent pas de ce que nous avons codé dans ces systèmes.