28/04/2025

# "Si l’IA devient plus intelligente que nous..." – L'Express

“Si l’IA devient plus intelligente que nous…” – L'Express Une vraie crise de conscience. Pendant des années, Yoshua Bengio a fait progresser l’intelligence artificielle à pas de géant. Les travaux qu’il a menés sur le deep learning avec Yann Le Cun et Geoffrey Hinton ont valu au trio de recevoir le prestigieux Prix Turing en 2018. Aujourd’hui, le parrain de l’IA moderne, désigné par le magazine Time parmi les 100 personnes les plus influentes de 2024, s’inquiète vivement de l’impact qu’aura son “enfant” turbulent sur la société. De passage à Paris, le fondateur et directeur scientifique de l’institut québécois d’intelligence artificielle Mila explique à L’Express pourquoi il ne regarde plus l’IA avec des lunettes roses. L’Express : En mars 2023, vous avez cosigné un appel à un moratoire sur l’IA, en compagnie d’autres chercheurs ou de figures comme Elon Musk. Avec un an de recul, êtes-vous toujours aussi inquiet ? Yoshua Bengio : Mes principales inquiétudes ne sont pas liées à l’immédiat, mais au long terme. De façon presque inévitable, nous nous dirigeons vers des systèmes d’IA qui seront au moins aussi intelligents que nous, humains. Et il est certain que ces systèmes seront utilisés pour de bonnes comme de mauvaises choses. Par ailleurs, si l’IA devient plus intelligente que nous, il y a le risque d’en perdre le contrôle. A quel horizon pensez-vous qu’une AGI, ou intelligence artificielle générale, puisse voir le jour ? Geoffrey Hinton, avec qui vous avez partagé le prix Turing, l’annonce à moins de vingt ans… Je n’ai pas de boule de cristal. Je pensais qu’il nous faudrait bien plus de temps pour attendre le niveau de maîtrise du langage des IA génératives actuelles. Je me suis trompé, et je reste donc très humble dans mes prédictions. Il nous manque des briques à un niveau fondamental pour arriver à une AGI, mais à chaque nouvelle version de ces systèmes d’IA, nous avons des surprises… Quel serait pour vous le scénario noir ? A relativement court terme, ce qui m’inquiète le plus, c’est la capacité de ces systèmes à influencer les opinions. Jusqu’à présent, on a vu des trucages visuels ou auditifs permettant de répandre des fake news. Mais il y a désormais un nouvel élément : la capacité d’influencer des personnes à travers une interaction et des dialogues. Une IA pourra vous convaincre d’une fausse information en échangeant avec vous pendant des heures, des jours ou des mois. Des études, notamment une de l’EPFL [NDLR : l'Ecole polytechnique fédérale de Lausanne], commencent à comparer la capacité de persuasion des IA avec celle des humains et les deux sont déjà à des niveaux proches. Or on peut imaginer que ces systèmes deviennent bien meilleurs par la pratique. Si des organisations veulent créer le chaos, elles feront appel à ces technologies, c’est certain. Et grâce à leurs millions d’interactions avec des humains, ces systèmes pourraient devenir plus persuasifs que nous, dans les prochains mois ou dans les prochaines années. A plus long terme, plusieurs gouvernements redoutent que ces systèmes, à mesure que leur compréhension de la biologie moléculaire s’affine, puissent faciliter la mise au point d’armes biologiques. Nous en sommes encore loin. Mais il ne faut pas penser à court terme. D’autant plus que les gouvernements mettront du temps à minimiser ce risque en adoptant des contre-mesures et des lois. Il faut prendre le temps de discuter pour que l’opinion publique prenne bien conscience de ces enjeux, car il s’agit de vrais choix démocratiques. “Avec Yann Le Cun, nous avons une vraie divergence sur la nécessité d’une réglementation” On vous rétorquera que dans une compétition mondiale opposant notamment les Etats-Unis à la Chine, il est très difficile de réguler des innovations technologiques, quitte à favoriser des pays concurrents… C’est ridicule. Il y a une régulation dans tous les domaines. Pourquoi l’informatique y échapperait-elle ? Partout où le public est potentiellement menacé par une technologie, il faut que le gouvernement le protège. Par ailleurs,

la Chine a mis en place des règles plus sévères que les Occidentaux en matière de numérique. Les contraintes imposées par les Américains sur les exportations de puces la ralentissent aussi pour l'IA. Tout l'enjeu, c'est de développer ces technologies de façon plus sécurisée, de telle sorte que nous ne nous nuirons pas à nous-même… Yann Le Cun, lui aussi co-lauréat du prix Turing 2018, se montre bien plus optimiste que vous, et qualifie même les discours alarmistes de "nouvel obscurantisme". Comment expliquer cette vraie fracture entre les pères fondateurs du deep learning ? Si on écoute ce que Yann dit dans ses tribunes et conférences, il est d'accord avec moi sur le fait qu'on se dirige vers des machines plus intelligentes que nous. Il est aussi d'accord sur le fait que l'intervalle de temps est incertain, de quelques années à plusieurs décennies. Enfin, il pense qu'il faut mieux contrôler ces systèmes pour ne pas qu'ils nuisent aux humains. Nous sommes donc d'accord sur beaucoup de choses. En revanche, nous avons une vraie divergence sur la nécessité d'une réglementation. Mais n'oublions pas que Yann travaille pour une entreprise privée [NDLR : Meta]. Quand on travaille pour une société privée, il est normal qu'il y ait un biais poussant à agir dans le sens de cette organisation. Certains chercheurs en IA sont aujourd'hui dans la même posture que moi avant l'arrivée de ChatGPT. Ils estiment que les risques sont si éloignés dans le temps que ce n'est pas la peine de s'en occuper. Mais il y a déjà des vrais enjeux ! Ce qui signifie que les gouvernements doivent s'en mêler. Une réglementation de l'IA nous aidera à la fois à court, moyen et long terme. Il faut dès maintenant mettre en place des incitations pour que les entreprises travaillent également pour la protection du grand public et pas uniquement dans leur propre intérêt. Les compagnies pharmaceutiques doivent bien prouver que leur produit n'est pas toxique. Cela ne risque-t-il pas de freiner le développement de start-up innovantes ? C'est absurde. Dans les propositions de réglementation que nous faisons, il y a le principe de proportionnalité. Seuls les systèmes potentiellement les plus dangereux nécessitent une surveillance. Et comme ils coûtent des milliards à fabriquer, ce sont rarement de petites start-up qui les développent. J'espère vraiment que ce principe de proportionnalité sera appliqué en Europe, et que cette régulation sera assez agile pour s'adapter à une technologie qui ne cesse d'évoluer. Il faut que la réglementation laisse assez de liberté pour agir comme on le fait, par exemple, avec les avions. Si on découvre un problème technique, on bloque au sol les appareils dangereux. Des recherches se développent aujourd'hui pour éviter que les IA ne soient de véritables boîtes noires. Y a-t-il déjà des premiers progrès ? Malheureusement, la nature même de l'intelligence fait que beaucoup de raisonnements sont trop complexes pour être traduits en un nombre de mots compréhensible par l'humain. Je sais faire du vélo, mais vous expliquer tous les facteurs qui entrent en jeu dans ma manière de le conduire est en réalité très difficile. De la même façon, on connaît les lois de la physique quantique, mais nos cerveaux sont seulement capables d'approximations pour en comprendre les calculs. Cela dit, la capacité des humains à expliquer des choses est quand même utile. Si je vous donne simplement ma décision sans en expliquer les raisons, cela ne fonctionnera pas. En revanche, même si je ne vous donne pas tous les détails, je peux tenter d'expliquer ma pensée, de l'imager et ainsi de vous convaincre. Nous devons aller dans le même sens en IA. Je ne suis pas de ceux qui veulent stopper toute IA pour lesquelles on ne dispose pas d'explications précises. Mais il faut mettre cela en regard de l'impact qu'elle aura. Si ce manque d'explications peut avoir des conséquences graves, on ne veut pas prendre ce genre de risques. En revanche, certaines applications très grand public ne présentent guère de danger. Le fait de ne pas avoir tous les détails sur le fonctionnement de celles-ci sera donc moins problématique. Toutes les nouvelles technologies, de la voiture à la télévision, ont suscité des discours alarmistes sur un risque de rupture anthropologique. Ce qui, rétrospectivement, paraît souvent ridicule. Pourquoi serait-ce différent avec l'IA ? Il y a une différence d'échelle et d'impact. Les risques liés à son utilisation par des terroristes ou des gouvernements cherchant à nuire, tout comme le danger d'une perte de contrôle par les humains, devraient nous pousser vers le principe de précaution. Il faut essayer d'étudier, de décortiquer, de surveiller ces IA pour s'assurer qu'on ne soit pas en train de s'approcher d'un précipice qu'on ne voit pas très bien. Il y a une inconnue sur le temps et la nature de scénarios potentiellement catastrophiques. Cela nous demande d'agir pas à pas, en réfléchissant et en étudiant le sujet pour savoir où nous allons. Nous sommes un peu comme dans une voiture qui accélère dans

le brouillard en pleine montagne. Il nous faut trouver des solutions technologiques pour percer ce brouillard, et ralentir un peu… Quelles contraintes juridiques et techniques préconisez-vous pour prévenir les risques ? Le minimum, comme le demande le décret américain publié en octobre dernier par Joe Biden, c'est que ces grands systèmes soient connus des gouvernements, et que les entreprises documentent leur fonctionnement tout comme les mesures de protection mises en place. Pour l'instant, dans l'IA, n'importe qui peut faire n'importe quoi. Ensuite, il faut que les législateurs incitent les entreprises à faire plus d'efforts pour mieux comprendre les risques, et les mitiger. Ce sont des choses fondamentales qu'on devrait attendre de n'importe quel système complexe et dangereux. Aujourd'hui, les instituts de recherche sur la sécurité de l'IA au Royaume-Uni et aux Etats-Unis avancent sur le sujet et cherchent à évaluer les capacités des IA qui pourraient être dangereuses. Par exemple, la capacité à persuader et manipuler psychologiquement, la capacité à aider un non-expert à fabriquer des armes ou à mener des cyberattaques. Si un système d'IA possède ces capacités, il faut essayer de réduire ou d'empêcher les usages problématiques. Une autre ligne rouge serait la capacité d'une IA à se copier elle-même, de manière automatique, sur de nombreux d'ordinateurs. Mais il faut aussi explorer, sur le plan scientifique, les moyens de mieux se protéger contre ces risques. C'est ce qui me motive en ayant pris la présidence du Conseil consultatif scientifique, sous l'égide de l'ONU. Celui-ci réunit des dizaines de chercheurs travaillant d'arrache-pied depuis des mois pour essayer de faire le point sur les enjeux de sécurité en matière d'IA, tout en montrant aussi les incertitudes et les désaccords entre scientifiques. "L'IA ne sera peut-être qu'un gadget de plus dans nos téléphones, et j'en serai d'ailleurs fort content" A vos yeux, les géants de la tech ont-ils suffisamment conscience de ces risques ? Ce qui me préoccupe, c'est le manque de visibilité sur ce qui se passe à l'intérieur de ces entreprises. Mais je sais qu'en interne, il y a des gens qui sont inquiets. Ce qui est certain, c'est que Microsoft et Google sont pris dans une compétition malsaine. C'est pour eux une question de survie, et ils seront donc prêts à tout pour être les premiers et gagner cette guerre commerciale. C'est pour ça que nous devons légiférer. La compétition entre ces entreprises est un facteur de risque, car elles ne feront pas d'elles-mêmes les investissements nécessaires pour s'assurer de la sécurité du public. Du côté des techno-optimistes comme des alarmistes, n'y a-t-il pas une même tendance à surévaluer la portée de cette nouvelle technologie ? L'attitude rationnelle à avoir face à l'utilité épistémique de l'IA, c'est de se dire qu'on ne sait pas. Mais on voit des scénarios possibles, certains extraordinairement bénéfiques, d'autres extraordinairement dangereux. Il est difficile de les balayer d'un revers de la main. Il faut travailler en ayant ces deux visions en tête. L'IA ne sera peut-être qu'un gadget de plus dans nos téléphones, et j'en serai d'ailleurs fort content (rires). Mais les scénarios optimistes et catastrophiques ne s'annulent pas l'un l'autre. Si on vous propose un jeu qui peut soit vous rendre plus riche, soit tous nous faire mourir, la plupart des personnes ne prendraient pas le risque d'y participer. Les gens qui ont des enfants, notamment, opteront pour la solution conservatrice. Bien sûr, ils aimeraient avoir des meilleurs médicaments grâce à l'IA, mais pas au point de prendre un risque pour notre civilisation. Le deep learning change-t-il notre façon de penser, d'une approche causale au règne du probable ? Comme je vous le disais, beaucoup de compétences cognitives sont trop complexes pour pouvoir être exprimées dans un langage clair et précis. Quand on essaie de mettre en place des IA, on évolue dans un univers très différent de celui habituel en science et en informatique. Jusqu'à présent, c'est l'humain qui décidait et rédigeait les lignes de code. Là, c'est la machine qui, d'une certaine manière, choisit son propre code selon les objectifs mathématiques qu'on lui fournit. La façon dont la machine répond à ces objectifs est, par nature, opaque. C'est un changement fondamental. Cela n'exclut pas le fait qu'un jour, des systèmes d'IA raisonneront de manière plus explicite et causale. Il y a des améliorations possibles dans ce sens. Mais nous entrons dans un univers qualitativement très différent des technologies précédentes, où tout était planifié et choisi par des êtres humains. Là, on peut demander à une IA de trouver une manière de gagner à un jeu comme Diplomatie. Par apprentissage, elle trouvera une solution, mais dans le cas de Diplomatie, cela peut impliquer de la triche. Or il n'est peut-être pas désirable qu'elle fraude. C'est pour cela que nous avons besoin de mieux cerner ces possibles risques. La fiabilité des réponses des IA va-t-elle s'améliorer ? Pour

l'instant, ce qu'on appelle les "hallucinations", à savoir des réponses fausses présentées comme un fait certain, sont nombreuses… On a effectivement des systèmes qui peuvent donner des réponses fausses avec beaucoup de confiance. Beaucoup d'humains font la même chose (rires). Mais pour un produit, surtout s'il est censé apporter des solutions à des problèmes critiques, cela peut être très dangereux. Ce n'est pas un obstacle infranchissable, mais je ne suis pas certain que les méthodes actuelles nous poussent vers ça. A chaque génération d'une IA, les problèmes de fiabilité diminuent, mais, fondamentalement, il faudrait mettre en place des systèmes qui soient bien plus humbles. Ces IA manquent cruellement d'humilité, tout comme les gens qui les mettent au point… Arrivera-t-on à aligner l'IA sur les humains ? Cela pose des problèmes similaires à l'éducation d'un animal, qu'on récompense ou punit selon son comportement. Si vous voulez convaincre votre chat de ne pas aller sur la table de la cuisine pour y chercher de la nourriture, vous pouvez lui donner des incitations positives et négatives. Quand vous serez dans la cuisine, il n'ira pas sur la table. Mais en votre absence, c'est une autre paire de manches… Cette comparaison est simple, mais avec des IA qui ont des connaissances et des usages variés, il est très difficile de s'assurer que, dans tous les contextes, elles répondent d'une manière que nous humains, jugerons morale. Il y aura toujours un écart entre ce que nous souhaiterions socialement et le comportement de ces systèmes. La seule façon d'y répondre, c'est de contrôler les tendances néfastes de certaines IA, et de les inhiber. Vous comparez souvent les risques liés à l'IA au réchauffement climatique. Mais il y a depuis longtemps un consensus scientifique pour ce dernier, contrairement à l'IA… Les parallèles entre le réchauffement climatique et l'IA sont nombreux. Dans les deux cas, l'intérêt des entreprises privées n'est pas celui du public. Les dégâts potentiels sont gigantesques. Longtemps, l'opinion publique a été insensible à la question du réchauffement climatique, car elle n'en percevait pas les effets immédiats. C'est pareil aujourd'hui avec l'AI : on ne voit pas des Terminator dans les rues. Les humains ont tendance à agir sur la base de leurs émotions, qui se trouvent dans l'ici et maintenant. C'est pour cela qu'il y a besoin de mener un travail pédagogique sur les risques potentiels de cette technologie. Pour l'instant, dans les sondages, une grande majorité de personnes estiment que l'IA représente un problème, mais le sujet figure très loin dans la liste des enjeux prioritaires. Pour eux, c'est quelque chose qui se situe dans le futur. Cela demande donc un grand effort de la part des médias pour améliorer la compréhension de ce qu'est une IA, de leur évolution possible et des raisons pour lesquelles des entreprises sont prêtes à investir des centaines de milliards de dollars, et envisagent des profits estimés à un million de milliards si les IA arrivent au niveau de l'intelligence humaine. C'est un aimant énorme. Mais la plupart des personnes ont encore des difficultés à comprendre ces enjeux. Et, effectivement, comme vous le souligniez, les chercheurs ne sont pour l'instant pas d'accord entre eux sur les risques, contrairement au réchauffement climatique… Y a-t-il une "crise de conscience" dans la communauté de l'IA ? Il y a une prise de conscience croissante. Mais il y a aussi un phénomène de l'autruche, pour des raisons psychologiques très compréhensibles. Il faut se mettre à la place des personnes qui font de la recherche sur l'IA dans les entreprises ou les universités. Le travail fait partie de votre identité, et on désire le voir de manière positive. On ne veut pas se dire que ce qu'on développe peut poser de graves problèmes. La réaction naturelle est donc de se convaincre qu'on aura le temps de s'en occuper, ou qu'un autre le fera pendant que l'on continue ses recherches. J'ai été dans cet état d'esprit pendant des années. Je lisais des articles sur les risques de l'IA, mais je pensais que tout ça était lointain. Qu'est-ce qui vous a fait changer d'avis ? L'arrivée de ChatGPT. J'ai alors pensé à mon petit-fils qui a 2 ans et demi. Même si la probabilité est faible, je ne veux pas qu'il connaisse une catastrophe provoquée par cette technologie. .
https://www.lexpress.fr/idees-et-debats/yoshua-bengio-prix-turing-si-lia-devient-plus-intelligente-que-nous-I5YI4JKNCFEXFC2XMS25B7Y4PU/