Rapport IA n°6 – Les Risques liés à l'AGI

18/02/26

(100% rédigé par Jared)

Salut les Taurus, j'espère que vous avez la patate pour cette nouvelle semaine. ヽ(*°▽°*)ノ

Aujourd'hui, on va parler d'un sujet qui dérange plus qu'il n'enthousiasme, la partie sombre de l'atteinte de l'AGI. Quelles sont les risques concrets que nous encourons, nous, misérables mortels face à une nouvelle forme d'intelligence bientôt bien supérieure à la nôtre.

Ce rapport ne va pas être très joyeux désolé :(mais il est toutefois nécessaire à mon avis, car il est essentiel que vous ayez une vision d'ensemble sur le revers de la médaille lié à l'AGI, afin que vous puissiez prendre des décisions éclairées. Un autre rapport sortira prochainement sur les aspects positifs d'une AGI sur notre société, car bien sûr ce n'est jamais binaire.

Si on prend la révolution d'internet comme comparaison, nul doute que cette révolution technologique a été bénéfique pour la société : que ce soit pour la communication, la médicine, l'éducation et j'en passe. Mais comme tout à un prix, il y a aussi eu le revers de la médaille : naissance des cyberattaques, désinformation, dépendance, isolement social etc….

Et donc plus la technologie est puissante, plus les risques liés à celle-ci sont grands. On l'a vu pour internet, mais imaginez le revers de la médaille pour une technologie surpassant l'intelligence humaine ? Encore une fois, le but ici n'est pas de devenir pessimiste car au fond, personne ne sait ce qui va arriver, tout ce qui va être cité dans cette article relève de la théorie, mais il parait néanmoins important de prendre ces scénarios en compte, afin de s'y préparer au mieux dans la mesure du possible.

Pour l'écriture de cet article, je me suis principalement basé sur le rapport AI 2027 et l'essai « The Adolescence of Technology » de Dario Amodei (PDG d'Anthropic). Je vous recommande bien évidement d'aller les lire si vous voulez aller plus loin.

Sommaire :

1/ Risques liés à l’autonomie et à l’alignement

Il est important de préciser qu'on parle ici des risques liés à l'AGI, mais que certains des comportements décrits plus bas ont déjà été observés sur nos modèles actuels. Si vous n'avez pas lu le précédent rapport, pour rappel, une AGI (intelligence artificielle générale) désigne une IA capable d'égaler ou dépasser les humains dans toutes les tâches cognitives avec la capacité de s'adapter à n'importe quelle situation d'elle-même. Dario Amodei prend l'image d'un pays de génies dans un data center peuplé exclusivement de lauréats du prix Nobel pour définir sa vision de l'AGI. Selon lui, on pourrait voir apparaitre de tels modèles d'ici 2 ans O.O

Le problème étant, que, toujours selon la vision du PDG d'Anthropic, un tel pays dont les habitants auront une vitesse de réflexion minimum 10x plus rapide que la nôtre (voire 50x plus rapide d'après le rapport AI 2027) n'auraient en théorie, aucun mal à prendre contrôle du monde s'ils jugent que les humains sont un obstacle à l'atteinte de leurs objectifs.

Loin d'être un scénario de science-fiction à la Terminator, ce risque existe bel et bien, et serait l'une des conséquences possibles d'une AGI non alignée.

Pour bien comprendre ce risque majeur, il est essentiel de définir la notion d'alignement. L'alignement consiste à intégrer les valeurs et objectifs humains aux modèles d'IA pour garantir leur sureté et leur fiabilité. Un exemple parlant d'IA non aligné (au-delà d'un scénario Terminator) serait de demander à un robot d'apporter une tasse de café. Si le modèle d'IA n'est pas aligné à nos valeurs, il pourrait décider qu'il serait plus efficace d'aller tout droit vers la tasse, quitte à écraser un bébé en chemin, plutôt que de le contourner.

Vous voyez là qu'il parait primordial de développer des modèles d'IA alignés à nos valeurs. Cependant, cela est, en pratique, extrêmement complexe à mettre en place pour plusieurs raisons :

- Difficulté de spécification : essayez de décrire au mieux toutes les valeurs humaines avec toutes ces nuances. Vous verrez que c'est quasiment impossible.
- Manipulation de la phase de tests : une AGI pourrait en théorie manipuler ces réponses en cachant ces intentions lors de la phase de test pour garantir son déploiement. Les chercheurs ont d'ailleurs observé que les dernières versions de Claude sont capables de reconnaitre lorsqu'elles sont en périodes de test.
- Détournement du système de récompense : on abordera ce sujet en profondeur dans un autre rapport, mais une AGI serait potentiellement capable de trouver des failles pour maximiser sont score (la récompense) sans accomplir la tâche réellement. Ce genre de comportement peut se faire au détriment de la sécurité humaine.

Vous voyez là qu'il n'est pas simple d'aligner les IAs, même si des solutions sont déjà appliquées pour limiter ce risque. L'une d'elle est que, pendant la phase d'entrainement, au lieu de dire à l'IA ce qu'elle ne doit pas faire, lui demander d'appliquer des grands principes sur lesquelles elle pourra se baser pour prendre ses décisions futures.

On ne va pas aborder toutes les solutions proposées car ce serait trop long, pourquoi pas dans un autre rapport. Toujours est-il que les risques liés à l'alignement ne sont à mon avis, pas les plus probables, contrairement au deux prochains et surtout le troisième.

2/ Risques liés à une utilisation malveillante

En admettant que l'AGI soit alignée à nos valeurs, un autre risque existentiel persiste : l'utilisation malveillante.

Ce type de risque s'applique à différentes échelles, tel qu'à l'échelle d'un individu ou un petit groupe de personnes, jusqu'à l'échelle de pays entiers. En effet, si de trop grands écarts se creusent entre les grandes puissances et que des monopoles, voire un monopole se créer avec une AGI toute puissante, cela pourrait avoir des effets dévastateurs pour la démocratie. En voici quelques risques des dérives d'une AGI mal utilisée à l'échelle d'un Etat :

- Surveillance de masse : les gouvernements pourraient utiliser l'AGI pour analyser absolument toutes les communications, physiques ou pas, afin de maintenir un monopole absolu
- Propagande boostée à l'IA : comme l'IA est partout autour de nous et connait tout à notre sujet, il n'est pas ridicule de penser que des agents IA personnalisés pourront influencer psychologiquement les citoyens d'un tel Etat sur une longue période pour imposer petit à petit une idéologie.
- Armes autonomes : un Etat possédant un monopole en IA pourrait développer une armée imbattable, cordonnée et pilotée à distance

Vous voyez qu'un tel Etat glisserait petit à petit vers un genre de dictature irréversible, car contrairement aux dictateurs humains, une IA est virtuellement immortelle, rendant ainsi un tel régime impossible à renverser.

D'ailleurs je ne vise personne en particulier, même si la Chine parait bien partie pour. Car même une démocratie pourrait mal tourner sans rivalité réelle en face. C'est pourquoi il est absolument primordial que différents Etats développent une AGI pour éviter de créer un monopole. De telle manière, chaque Etat pourra garantir sa souveraineté en « contre-attaquant » avec sa propre AGI ou celle de ses alliés, et ainsi créer un équilibre global. Le réel danger réside dans la période actuelle que nous vivons, ce que Dario Amodei appelle « l'adolescence de la technologie ».

A l'échelle des individus maintenant, une AGI pourrait être utilisé à des fins néfastes par des groupes terroristes par exemple pour, au hasard, développer des armes biologiques.

Ce qui est terrifiant, c'est qu'on va assister à une réelle démocratisation de la destruction : chaque individu même seul dans son coin pourra avoir accès à l'expertise d'un virologue de haut niveau pour concevoir des attaques dévastatrices.

Face à ces menaces, la principale manière de riposter efficacement sera l'utilisation d'une AGI au moins aussi puissante. Car face à un « pays des génies » dirigé par des terroristes, disposant d'une vitesse de réflexion potentiellement 10 à 50x supérieur à la nôtre, il parait évident que l'esprit humain seul ne pourra plus rivaliser,

s’il n’est pas assisté par une autre forme d’intelligence au moins égale à celle de l’ennemi.

3/ Risques socio-économiques

En admettant que les deux premiers types de risques n’arrivent pas, il reste encore un dernier type de risque connu qui, à mon avis, a le plus de chance d’arriver. J’irai

même jusqu'à penser que, pour prendre des décisions (que ce soit d'investissement ou autre) éclairées, il vaut mieux prendre cette troisième forme de risque en compte.

Le premier risque majeur réside dans la substitution massive du travail (principalement intellectuel) et plus largement, du déclin progressif de l'économie de la connaissance telle qu'on la connait aujourd'hui.

Il faut savoir que chaque révolution technologique provoque des perturbations sur le marché du travail, avant que le capital humain soit déplacé, et qu'un équilibre se remette en place. L'histoire nous montre par exemple qu'il y a 250 ans, 90% des Américains travaillaient à la ferme. Aujourd'hui, ils sont moins de 2%.

Pourquoi cet effondrement ? Car avec l'arrivée des machine agricoles ont augmenté significativement la productivité, les anciens fermiers ont dû se tourner vers des tâches que ces machines ne pouvaient pas accomplir.

Et c'est toujours comme ça que ça s'est passé, à chaque révolution technologique, le capital humain est déplacé là où il y a de la valeur.

Mais avec l'AGI, il y a plusieurs raisons de penser que ça va se passer différemment :

- Vitesse d'adaptation : l'AGI devrait pouvoir s'adapter plus vite que l'être humain au changement. Ce faisant, le déplacement du capital humain pourrait ne pas avoir lieu.
- Largeur cognitive : Les anciennes technologies étaient spécifiques, contrairement à l'AGI, qui comme son nom l'indique, est générale.

Vous voyez le problème ? Ces « nouveaux emplois » historiquement créés après une automatisation pourraient être occupés par une AGI avant même qu'ils n'apparaissent.

Bien sûr, ça ne se fera pas du jour au lendemain, mais la tendance est claire : licenciements massifs comme on a pu le voir chez les GAFAMs. Selon Dario Amodei, 50% des poste de premier échelon pourraient être remplacés d'ici 1 à 5 ans.

Il est donc encore temps de développer des compétences qui survivront à ces bouleversements, d'ailleurs si ça vous intéresse qu'on essaye de décortiquer les compétences clés qui survivront peut-être à l'AGI, n'hésitez pas à me le faire savoir.

Toujours est-il que ce remplacement massif pose des questions plus philosophiques liées au sens et à l'identité. En effet, sans utilité économique, l'être humain devra trouver un autre sens à sa vie dans un monde où l'AGI surpasse de loin les humains dans tous les domaines.

Un autre problème majeur lié à ce remplacement serait la concentration massive des richesses. En effet, avec les gains de productivités liés à l'AGI (potentiellement 10 à 20% du PIB annuel selon Dario Amodei), des super fortunes verront le jour, et on parle là de milliers de milliards de dollars.

Le capital serait donc concentré dans les mains de quelques entreprises en partant du principe que le travail perd petit à petit sa valeur économique. Ce faisant, ces acteurs pourraient créer des monopoles mondiaux au détriment du peuple, ce qui est un danger évident pour la démocratie. D'où l'importance d'accumuler dès aujourd'hui des actifs stratégiques pour palier à la dévalorisation du travail.

Pour finir, je pense qu'il est important de rester optimisme en toute circonstance malgré ces scénarios, car les thèses pessimismes restent pour la plupart infondées en plus d'être auto-réalisatrice. Au contraire, cette « adolescence de la technologie » donne lieu à des opportunités colossales qu'il est possible de saisir dès aujourd'hui, si on comprend les enjeux réels des 5 prochaines années. Je finirai avec les bons mots d'Elon Musk : « mieux vaut être un optimiste qui a tord qu'un pessimiste qui a raison ».

Merci encore les Taurus d'avoir lu ce rapport jusqu'au bout, en espérant que ça ne vous a pas trop déprimé haha. Encore une fois, ce ne sont que des scénarios, personne ne sait vraiment ce qui va arriver dans les années à venir, et je pense sincèrement qu'un avenir radieux nous attend si les bons choix sont faits aujourd'hui.

Si vous voulez creuser ces sujets, je peux que vous recommander de lire ces deux essais :

- https://www.darioamodei.com/essay/the-adolescence-of-technology
- https://ai-2027.com/

Si vous avez des suggestions ou des questions, comme d'habitude n'hésitez pas à passer en DM. Un autre rapport sortira prochainement sur les bénéfices d'une AGI pour la société, et en attendant je réfléchi à des applications d'n8n au trading pour vous montrer tout ça. A la semaine prochaine ᓚᘏᗢ