Par Eric Stefanello,
Polytechnicien, spécialiste de l'IA et de la sécurité des logiciels et des systèmes complexes.

SÉCURITÉ DES IA : UN IMPÉRATIF FACE À L'INCONNU

Depuis quelques années, différents scientifiques et ingénieurs commencent à évoquer les problèmes que posent les IA neuronales en matière de sécurité des personnes. Cette question devient plus urgente alors qu'arrivent des IA interagissant directement dans et avec le monde physique.

Bien qu'abreuvé au quotidien par des informations sur les IA, le citoyen n'a pas une compréhension claire de ce que sont vraiment ces « machines », alors même qu'elles se déploient de plus en plus largement dans nos entreprises et dans nos vies.

Leur fonctionnement représente un changement de paradigme épistémique total car il utilise l'approche inductive plutôt que l'approche hypothético-déductive.

Les IA neuronales utilisent en effet des approches purement inductives. A ce titre, elles n'ont aucune représentation du monde sous-jacente, quel que soit le sujet sur lequel elles travaillent. Elles ne sont pas construites à partir de représentations symboliques du monde mais à partir d'une approche statistique massive de corrélation de grandes masses de données.

Plutôt qu'intelligences artificielles, on devrait donc les appeler des « corrélateurs massifs de données ».

C'est ainsi que dans les LLM (*Large Languages Models*) le langage, les mots sont représentés par des vecteurs de 300 dimensions. Chacun de ces mots/vecteurs possède une signification qui n'est absolument pas symbolique mais qui est basée sur la proximité statistique avec les autres nombres des autres mots ayant un sens voisin ou apparenté.

Quand cette vectorisation du langage est bien faite, on obtient une représentation calculable au sens mathématique. C'est ainsi que si vous prenez le vecteur roi et que vous lui retranchez le vecteur homme et que vous lui ajoutez le vecteur femme, vous obtenez alors un vecteur qui est extrêmement proche d'un vecteur reine. Extrêmement proche au point qu'il n'y a pas d'autre vecteur aussi proche du vecteur reine. Vous pouvez donc inférer que : roi - homme + femme = reine.

Cette calculabilité est LE pilier du fonctionnement de tous les LLM. Elle permet de faire des calculs sur les mots, de les trier, de les distinguer, de les classer, de les contextualiser, de les interpréter.

C'est là qu'intervient la phase d'apprentissage : on va faire ingurgiter au réseau neuronal des masses de données qui vont lui permettre peu à peu de corréler celles-ci et d'être capable d'inférer, après une suite de mots / vecteurs calculables, quel est le mot suivant qui peut advenir de façon la plus probable.

On construit ces réseaux de neurones en les dotant de paramètres qui vont intervenir dans chaque entrée et dans chaque sortie de neurone. C'est l'apprentissage qui va permettre à la machine de régler peu à peu chacun de ces paramètres jusqu'à ce que ses performances correspondent à ce qui est attendu.

Le programme n'est pas maitrisé, ni conçu par l'homme, c'est la machine qui s'autoprogramme grâce aux données qui lui sont fournies.

Le « programme » est le résultat de l'apprentissage, pas celui d'un travail humain. Corollaire : on ne sait pas vraiment ce qu'il y a à l'intérieur des IA, pas plus que l'on maîtrise leur comportement, et surtout leur comportement est non déterministe : il ne peut être, ni prédit, ni analysé précisément a posteriori.

Donc, si intelligence il y a, un jour, peut-être, celle-ci sera totalement, radicalement et définitivement différente de l'intelligence humaine. Les IA neuronales seront probablement omniscientes mais manqueront des formes d'intelligence humaines qui sont propres à notre biologie : intelligences relationnelle, psychomotrice, musicale et surtout connaissance de soi puisque cette dernière est liée à la conscience.

Il en ressort que les IA posent des problèmes de sécurité radicalement nouveaux : pourrons-nous un jour faire confiance à ces machines, au point de les laisser piloter des pelleteuses dans la rue, des avions remplis de passagers, des usines de retraitement des eaux, des réseaux électriques… ?

Avec les réseaux neuronaux, l'humain a inventé des programmes qui « s'auto-modifient » et apprennent tous seuls, sans notre concours si ce n'est celui de nos données ou des données qu'elles acquièrent autrement. On sait comment ils apprennent, mais on ne connait pas le contenu de tous leurs nouveaux apprentissages. C'est seulement lorsque l'IA agit que l'on découvre ses nouvelles capacités. Ce sont des programmes qui s'auto-conçoivent en quelque sorte, des bébés programmes qui deviennent de façon autonome adultes, par l'auto-apprentissage.

Sur le plan de la prospective on peut, sans prendre de risques, annoncer l'arrivée massive des IA neuronales dans le monde réel dans les 10 ans à venir. Ces IA agissant directement dans le monde réel vont poser des problèmes de sécurité radicalement nouveaux.

En effet, les systèmes logiciels les plus complexes utilisés dans les systèmes physiques sont jusqu'à aujourd'hui totalement déterministes dans leur comportement : pilotage des avions, contrôle des centrales nucléaires…

Aucun système informatique de ce type n'a jamais eu de panne massive. La multiplication des calculateurs et la certification des logiciels qui tournent dessus crée une « *safety by design* ».

Dans le monde des IA neuronales d'aujourd'hui et de demain, rien de tout cela. Nous sommes face à des programmes dont le comportement recèle intrinsèquement un aspect indéterminé, ne serait-ce que parce que l'apprentissage modifie le « programme » en permanence.

Autre facteur de risque majeur, si les mathématiques qui permettent de fabriquer ces réseaux neuronaux sont simples, nous n'avons pas les mathématiques qui permettent de décrire ou de comprendre leur fonctionnement. Oui, vous avez bien lu : les mathématiques décrivant le fonctionnement des IA n'existent pas encore.

Il faut savoir que les cathédrales informatiques que sont ces grands LLM comme Chat GPT avec 1800 trillions de paramètres et plus de 150 couches de neurones relèvent plus d'une construction de geek et de hackeurs que d'un ordonnancement organisé de maîtres maçons sous la direction d'un architecte avisé. Ces édifices se sont construits peu à peu avec des jeux d'essais et d'erreurs avec une constante : les résultats obtenus allaient toujours au-delà des espérances initiales et surprenaient les concepteurs.

Or il n'existe pas de possibilité de contrôle des IA si nous n'avons pas de modélisation mathématique précise de celles-ci. Cette question est fondamentale.

Depuis 18 mois les créations d'organismes étatiques dédiés à la sûreté des IA se multiplient dans le monde. Les USA en novembre 2023, puis le Japon, la Corée, Singapour, le Canada, la Chine, l'Argentine en 2024. L'UE a créé un bureau dédié fin 2024 et la France vient de créer l'INESIA (Institut national pour l'évaluation et la sécurité de l'intelligence artificielle) qui n'est pas une entité dédiée mais une fédération sous l'égide du SGDSN (Secrétariat général de la défense et de la sûreté nationale), des différents organismes impliqués dans les questions de sécurité et de sûreté des IA.

Lors du sommet sur la sécurité de l'IA en novembre 2023 à Bletchley Park et à l'initiative du gouvernement britannique, a été constitué un panel de 96 experts internationaux provenant de 30 pays, ainsi que de représentants d'organisations internationales telles que les Nations Unies, l'Union européenne et l'Organisation de coopération et de développement économiques (OCDE). Le secrétariat du panel est assuré par le gouvernement britannique, et le professeur Yoshua Bengio en assume la présidence pour l'année 2025.

On pourrait considérer qu'il s'agit d'une sorte de « GIEC de l'IA ».

Ce panel vient de sortir en janvier 2025 son rapport : « International AI safety report » disponible en ligne. Après quelques rappels sur le fonctionnement des IA, ce rapport, passé largement inaperçu dans la presse ces dernières semaines, malgré le sommet IA de Paris, apporte des éléments édifiants en matière d'analyse des risques associés à l'IA.

Constatant que les performances des dernières versions de ChatGPT arrivent maintenant au niveau de celles des meilleurs étudiants en PhD, le rapport pointe que notre compréhension limitée des implications des risques de l'IA constitue un « challenge majeur pour les décideurs politiques : ils doivent dès à présent peser les bénéfices et les risques des avancées imminentes de l'IA, sans avoir pour autant de larges résultats scientifiques disponibles ». Il renforce le message en affirmant que les dernières tendances des performances des IA doivent constituer une « priorité urgente pour la recherche fondamentale sur la sûreté des IA dans les mois à venir ».

Le rapport poursuit en affichant qu'une mitigation préemptive des risques basée sur des preuves incomplètes ou limitées pourraient être ineffective ou inutile. D'un autre côté, attendre des preuves plus fortes des risques imminents qui sont posés, pourrait laisser la société prise au dépourvu et même interdire toute mitigation ultérieure.

La conclusion du rapport est on ne peut plus claire : l'avenir des IA générales est incertain avec un large éventail de trajectoires possibles dans le futur proche entrainant à la fois des conséquences très positives et très négatives.

Ces IA générales seront capables d'être à la fois :

- Des ingénieurs : en analysant un problème, en le modélisant, en l'optimisant. Elles pourront analyser et définir les tâches nécessaires pour arriver au but.
- Des scientifiques : en inférant des modèles à partir d'observations empiriques et de raisonnements contrefactuels (événements qui ne se sont pas réalisés mais auraient pu l'être sous certaines conditions).
- Des ouvriers, des soldats, des artisans : en dirigeant et manipulant des objets physiques (robots ou autres) qui utiliseront eux-mêmes d'autres outils pour remplir leurs tâches.
- Des artistes : en créant *ex nihilo* de nouvelles formes d'expressions artistiques dans tous les domaines.

Elles ouvriront le champ aux machines concevant et fabriquant d'autres machines, et créeront de nouvelles connaissances, de nouveaux savoirs, inconnus des hommes.

Bien entendu, au milieu de ces développements, il y a aura des choses fantastiques : de nouvelles compréhensions de l'univers, de la biologie, de nouveaux médicaments, moins de travaux harassants, plus de loisirs (pour peu que l'on mette en place les filets sociaux qui seront indispensables et qu'une vraie redistribution des richesses soit mise en place).

Mais il y aura aussi des risques importants. On voit déjà émerger des comportements émergents d'autoprotection dans les IA actuelles comme l'affirme Yoshua Bengio qui prétend avoir des informations très confidentielles sur ces sujets gênants pour les tycoons de l'IA.

Comme on peut faire confiance à l'infinie cupidité des milliardaires du digital et au désir de suprématie des autocrates, on peut s'attendre à ce que tout ce qui se mettrait en travers d'un développement sans contrainte des capacités des IA générales soit dûment combattu par tous les moyens possibles.

Les IA se développent très rapidement : iront-elles plus vite que nos capacités à les encadrer et à nous mettre en sécurité face à elles ?

Il faut donc absolument règlementer *ex ante*, ne pas attendre que les problèmes arrivent, inverser cette fatalité qui a fait que depuis 50 ans, la règlementation du digital a toujours eu un train de retard par rapport aux développements techniques du domaine.

Lien vers la version longue de l'article :
https://www.academia.edu/128168155/IA_EST_vdef_fev