Réponse à Yann Lecun :
l'IA n'a pas besoin de vouloir dominer le monde pour nous nuire

Dans un entretien à France Info, Yann Lecun, célèbre fondateur du laboratoire de recherche en IA de Facebook, explique que l'intelligence artificielle ne s'accompagne pas d'une volonté de domination, à la différence des hommes. Voilà pourquoi il se trompe.

Je viens d'écouter sur France Info un entretien de 15 minutes avec Yann Lecun, fondateur du laboratoire de recherche en IA de Facebook (FAIR).

Yann Lecun cherche à faire retomber la hype qui entoure l'IA et le deep learning en particulier. Il explique notamment comment l'IA a besoin d'une centaine d'heures pour atteindre aux jeux Atari le niveau qu'un humain atteindrait en 15 minutes.

Dans un sens il n'a pas tort de vouloir verser un peu d'eau froide sur le sujet.

Mais bon, ce qu'il ne dit pas sur cet exemple par exemple, c'est qu'en quelques heures de plus, l'IA atteint un niveau surhumain. Ce que fait AlphaZero notamment est particulièrement bluffant : pouvoir battre le meilleur logiciel aux échecs après seulement deux heures d'apprentissage.

C'est le travers classique dans lequel beaucoup tombent, confondre capacité et manière. Oui notre cerveau est une merveille qu'on ne comprend guère, l'IA est encore très bête à bien des égards. Mais l'IA nous dépasse déjà sur tant de sujets, qu'importe si elle le fait à sa manière, moins élégamment que nous ? Une IA qui a besoin de temps pour se former, qui consomme énormément d'énergie, mais qui peut faire mieux que nous in fine est vouée à bouleverser notre monde d'une façon ou d'une autre.

L'intelligence est avant tout à mon avis la capacité à résoudre des problèmes. D'un point de vue pragmatique, qu'importe encore une fois si l'IA ne s'y prend pas comme nous, si elle fait mieux que nous ?

Yann Lecun joue trop les puristes je trouve. Sachant que des IA de plus en plus capables vont apparaître compte tenu des investissements en jeu.

Yann Lecun dit aussi que l'IA, à la différence des hommes, ne s'accompagnera pas nécessairement d'une volonté de domination. Mais c'est encore une objection classique, c'est surprenant que ça vienne de lui. Max Tegmark la traite dans son livre Life 3.0 : le problème, ce n'est pas qu'une IA veuille dominer le monde, ou soit méchante, "evil" par nature. Le problème, très prosaïquement, c'est :

- · i. : une IA ultra capable dont les objectifs ne seraient pas PARFAITEMENT alignés avec les nôtres;
- · et/ou ii. une IA capable mais pas assez robuste.

i. : on aurait bien du mal à définir nos valeurs, difficile de penser à une proposition qui ne serait : ni trop floue et sujette à interprétation diverse et dangereuse pour nous (on veut être heureux : ok, votre cerveau sera mis dans un bocal shooté à la dopamine) ; ni trop précise et facilement détournable via des cas particuliers tordus. Une IA très capable, quels que soient les objectifs finaux qu'on pourrait lui assigner, devrait développer assez naturellement ces objectifs intermédiaires :

1. se préserver, empêcher qu'on la débranche ;
2. maximiser son accès aux ressources et à l'énergie ;
3. développer sa curiosité et rechercher "la vérité" pour en déduire la meilleure compréhension du monde possible.

Chacun des points peut conduire à notre perte si nos valeurs ne sont pas PARFAITEMENT alignées avec celles de l'IA. Un enjeu lié : faire comprendre, adopter, retenir nos valeurs à l'IA. La fenêtre de temps où l'IA est assez capable pour comprendre nos valeurs, mais pas trop pour ne pas comprendre qu'elles sont peut-être futiles, sera peut-être très brève. Une IA trop capable pourrait comprendre que nos valeurs sont insignifiantes, de la même façon

que si l'intelligence humaine avait été créée par les fourmis avec pour objectif de protéger les fourmilières, on serait vite passé à autre chose par ennui, en massacrant des fourmis en chemin, non pas méthodiquement par objectif, mais en simples dommages collatéraux de la construction de nos infrastructures...

L'enjeu i. se résume à "Build the right system"

Une IA capable mais pleine de bugs de faille pourraient être détournée par de méchants humains, ou nous tuer par accident.

L'enjeu ii. se résume à "Build the system right"

C'est le cœur des recherches en AI Safety, avec comme 3ème enjeu lié :

- "measure" (interpretability, comprendre ce qu'il se passe dans la black box, bonne chance),
- et "control" (interruptability, peut-on concevoir un off switch button ?).

Pour ces débats, il vaut mieux suivre Deepmind et OpenAI.

En parlant de capacité des IA, voilà un dernier développement, les images ci-dessous ont été générées par l'IA, ce qui se fait de mieux à date en synthèse d'image, à base de GAN (Generative Adversial Network : une IA génératrice pond une image pour essayer de tromper une autre IA discriminante).

Deepmind a essayé de "scaler" cette démarche connue en utilisant plus d'images et plus de forces de calcul. Certains diront "The good news is that AI can now give you a more believable image of a plate of spaghetti. The bad news is that it used roughly enough energy to power Cleveland for the afternoon." Certes cela consomme beaucoup d'énergie mais c'est bluffant, et cela ne peut que s'améliorer. On se rapproche du jour où l'IA saura générer un épisode de Game of Thrones en une seconde.

Thomas Jestin, fondateur de KRDS et OhMyBot