L’IA n’est pas sans danger… Et c’est Google qui le dit

Pour Google, les inquiétudes entourant l’intelligence artificielle ne sont pas infondées. Dans un article de recherche, trois chercheurs de la firme, qui investit massivement dans ces technologies, associés à des universitaires et à un expert d’OpenAI (un institut de recherche cofondé par Elon Musk pour travailler sur la sécurité de l’intelligence artificielle) pointent cinq problèmes sur lesquels il faudra se pencher pour rendre les machines intelligentes plus sûres. Et éviter les accidents découlant de leur utilisation.

L’objectif affiché par Google ? Sortir des imprécations entre idolâtres et détracteurs de l’IA, pour entrer dans le concret. « *Alors que les risques potentiels de l’IA ont reçu une large attention de la part du public, les discussions autour de ce sujet sont restées très théoriques et basées sur des spéculations* », estime Chris Olah, l’un des auteurs de ‘Concrete Problems in AI Safety’, le travail de recherche publié hier. Et de militer pour le développement « *d’approches pratiques d’ingénierie de systèmes d’IA opérant de façon sûre et fiable* ».

Robots destructeurs ou tricheurs

Dans leur étude, les chercheurs mettent en lumière cinq problèmes qui, selon eux, se révéleront très importants à mesure que l’IA se répand dans la société. « *Des problèmes mineurs aujourd’hui, mais qu’il est important de régler pour les systèmes futurs* », juge Chris Olah. Premier écueil selon les chercheurs : les effets de bord. Ce qui revient à se demander quels garde-fous seront posés pour éviter que les machines intelligentes ne viennent perturber leur environnement afin de mener à bien leurs missions. Et Chris Olah de donner l’exemple d’un robot-nettoyeur qui renverse un vase pour être plus efficace dans sa tâche. Second obstacle, du même ordre : comment éviter que les machines ne ‘trichent’ pour mener à bien leurs missions. Par exemple, le même robot-nettoyeur pourrait, de lui-même, découvrir qu’il remplit les objectifs de son programme simplement en cachant les saletés plutôt qu’en les ramassant.

Au-delà des conséquences bénignes de l’exemple choisi par le chercheur, ces deux champs de recherche masquent des enjeux essentiels de la sécurité de l’IA. Et sont à l’origine des critiques de ses détracteurs, qui, pour certains, y voient une menace pour l’Humanité.

Les dangers de l’inexpérience

Autre difficulté répertoriée par le chercheur de Google : l’exploration de l’environnement et ses conséquences possibles. En partant toujours de l’exemple du robot-nettoyeur, le chercheur explique que l’IA embarquée dans ce dernier va probablement se lancer dans des stratégies basées sur l’utilisation d’une éponge. Mais que se passera-t-il quand le robot en question devra nettoyer une prise électrique ? Le quatrième problème répertorié par les chercheurs est assez proche et consiste à se demander comment s’assurer du comportement sûr d’un robot dans un environnement différent de celui où il s’est entraîné. Autant de cas où le robot peut provoquer, par

inexpérience, des accidents ou entraîner son auto-destruction.

Enfin, les chercheurs s'interrogent sur la façon dont l'IA pourrait respecter des aspects de sa programmation qui sont trop ennuyeux ou coûteux à évaluer fréquemment. Imaginons par exemple que le robot-nettoyeur doive trier des objets, en mettant à la poubelle ceux sans utilité et en rangeant ceux appartenant à des humains. Comment mènera-t-il à bien sa tâche sans avoir à solliciter en permanence ses propriétaires ?

Notons que le groupe DeepMind de Google, situé à Londres, vient également de publier un article de recherche sur la sécurité des machines intelligentes dans les environnements complexes. Cette recherche se penchait sur une autre question essentielle : comment éviter que les IA n'apprennent à shunter les sécurités mises en place par les humains pour interrompre leur fonctionnement en cas de comportement dangereux ?

A lire aussi :

Cédric Villani, « Plus il y aura d'IA, plus il y aura besoin des mathématiciens »

Les robots vont détruire et créer des millions d'emplois

Bill Gates se méfie des progrès de l'intelligence artificielle

Crédit photo : John Williams RUS / Shutterstock