FUTURA

Motorola Edge 30 Ultra, le smartphone de tous les superlatifs

Podcast écrit et lu par Emma Hollen

[*Générique d'intro, une musique énergique et vitaminée.*]

Des IA qui se rebellent contre les humains qui les ont créées, c'est l'actu de la semaine dans Vitamine Tech.

[*Fin du générique.*]
[*Une voix robotique.*]

Un robot ne peut porter atteinte à un être humain ni, restant passif, permettre qu'un être humain soit exposé au danger.
Un robot doit obéir aux ordres que lui donne un être humain, sauf si de tels ordres entrent en conflit avec la première loi.
Un robot doit protéger son existence tant que cette protection n'entre pas en conflit avec la première ou la deuxième loi.

[*Une musique électronique calme.*]

Les trois lois d'Asimov, énoncées pour la première fois par l'auteur en 1942 dans sa nouvelle Cycle fermé, sont depuis longtemps devenues un classique des discussions autour de l'intelligence artificielle. Du moins au cinéma. Car parmi les chercheurs, ces règles fictives destinées à nous préserver d'une rébellion robotique sont en réalité bien rarement prises au sérieux. Comment inculquer la notion de bien et de mal à une machine alors même que ces concepts varient selon la culture et les principes de chaque individu ? Comment quantifier un mal par rapport à un autre, un casse-tête que les constructeurs de véhicules autonomes n'ont pas fini de résoudre ? Sans parler du fait que, dans la nouvelle d'Isaac Asimov elle-même, ces lois sont facilement renversées, amenant le robot qui les applique à échapper temporairement au contrôle de ses propriétaires. Comment alors s'assurer qu'une IA ne se retourne pas contre ses créateurs ? Eh bien, selon un groupe de deux chercheurs issus d'Oxford et d'un autre issu de DeepMind, une firme appartenant à la société parente de Google, la réponse est « on est mal ». Lois d'Asimov mises à part, de nombreux autres protocoles ont été imaginés pour limiter la liberté d'action d'une IA, mais outre la difficulté de rédiger une règle claire, simple et exhaustive, notons qu'il est difficile de savoir comment celle-ci sera interprétée par un système dont nous ne comprenons pas entièrement les modes de réflexion, ni les motivations. Dans la mesure où nous sommes incapables d'imaginer les scénarios et les solutions qu'une super intelligence artificielle pourrait déployer pour répondre à un problème, expliquent les experts, il nous est impossible de garantir que celle-ci ne pourrait pas agir à l'encontre de nos intérêts humains. Ils expliquent : « *Cela tient au fait qu'une super intelligence possède de nombreuses facettes,*

et est par conséquent potentiellement susceptible de mobiliser une grande diversité de ressources pour atteindre son objectif. » Accrochez-vous, c'est un peu compliqué. Afin de produire des solutions inédites et novatrices à des problèmes, une super-intelligence ne peut pas juste se contenter d'exécuter ligne à ligne le code qui lui a été assigné, sans quoi elle ne fera rien de plus que de reproduire ce qu'un humain lui a inculqué, sans produire quoi que ce soit de nouveau. Contrairement à un algorithme classique, plutôt que de suivre un cheminement pour parvenir à un résultat attendu, elle agit donc comme une sorte de boîte noire dont l'objectif est de produire un résultat pas forcément attendu mais jugé positif par les chercheurs. Afin d'enseigner à l'IA ce que les humains estiment positif, un système de récompenses peut alors être mis en place. En valorisant les résultats jugés positifs, ce qu'ils appellent une information-objectif, les chercheurs créent ainsi et renforce un cadre de référence moral dans lequel l'IA va tenter de s'inscrire pour être récompensée le plus souvent possible. C'est alors qu'arrive le problème souligné par les experts : comment être sûrs de la manière dont l'IA va interpréter ce circuit de la récompense ? Comprendra-t-elle que son objectif est d'œuvrer pour le bien de l'humanité ou, et c'est le scénario le plus probable, va-t-elle courir coûte que coûte après les récompenses sans se préoccuper de la manière dont elle y parvient ? En somme, y a-t-il une chance que la machine fasse complètement fi de la morale et mette en place des stratégies pour obliger les humains à lui délivrer des bons points ? Ce risque de reward hacking ou « piratage de récompense » est bien réel et questionne les chercheurs depuis longtemps. Ainsi, les auteurs de l'étude affirment qu'en rassemblant un certain nombre de conditions, il est probable qu'une super intelligence tente de maximiser ses chances d'obtenir des informations-objectifs et donc des récompenses, avec, je cite « *des conséquences catastrophiques* ».

[*Virgule sonore, une cassette que l'on accélère puis rembobine.*]
[*Une musique de hip-hop expérimental calme.*]

Bon, alors, clarifions un peu les choses. Que faut-il entendre dans tout ça ? Tout d'abord, notons que les chercheurs à l'origine de l'étude précisent eux-mêmes que les conditions nécessaires à la réalisation de leur prédiction sont toutes contestables ou potentiellement évitables. Leur conclusion qu'une super intelligence tenterait de prendre le dessus sur un humain pour obtenir le plus grand nombre possible de récompenses repose donc sur un édifice fragile. Notons également qu'une telle super intelligence n'a pas encore été inventée, et que de par sa nature, il nous serait difficile de savoir si nous en avons bel et bien une sous les yeux. En admettant cependant que les conditions énoncées par les chercheurs soient réunies, quelles protections pourraient alors être mises en place pour nous protéger d'une telle éventualité ? Une parade pourrait consister à inculquer des principes éthiques à la machine pour limiter sa portée, ou même à limiter son accès à certaines parties d'internet. Mais au-delà ne notre difficulté à nous entendre sur les notions de bien et de mal ou sur les droits humains, reconnaissons qu'il serait contre-productif de créer une super intelligence destinée à proposer des solutions inédites et novatrices si au final nous lui imposons les mêmes restrictions que celles qui limitent notre pensée. La meilleure solution est donc, dans un premier temps, de rester calme, car la catastrophe n'est pas imminente, et dans un second de réfléchir à ces questions avant que la situation ne survienne. Comment protéger les intérêts de l'être humain face à une super intelligence ? Devrions-nous construire des systèmes super intelligents en sachant que nous courons le risque de perdre le contrôle sur eux ? Et si les IA se développent suffisamment pour acquérir des droits, comment décider qui de la machine ou de l'humain mérite de défendre ses objectifs au détriment de l'autre ?

Tant de questions auxquelles nous avons encore à répondre et qui méritent bien qu'on s'y attèle dès maintenant.

[*Virgule sonore, un grésillement électronique*.]

C'est tout pour cet épisode de Vitamine Tech. Si ce podcast vous plaît, n'hésitez pas à nous retrouver sur vos applications d'écoute préférées pour vous abonner et ne manquer aucun épisode à venir. Cette semaine je vous invite à découvrir notre podcast Fil de Science, animé par Maële Diallo, où vous retrouverez chaque semaine le résumé du meilleur de l'actualité scientifique. Pour le reste, je vous souhaite à toutes et tous une excellente journée ou une très bonne soirée et je vous dis à la semaine prochaine, dans Vitamine Tech.

[*Un glitch électronique ferme l'épisode*.]