UNE ETUDE POINTE LES POSSIBLES EFFETS PERVERS ET DANGERS DE L'INTELLIGENCE ARTIFICIELLE

Par Sébastien Gavois

Phishing parfaitement calibré pour une cible, malware avec une incroyable capacité d'adaptation, robot détourné de sa fonction première pour identifier et détruire une cible, système prédictif de perturbation civile et « fake news ». Voici quelques exemples des dérives possibles de l'intelligence artificielle, mises en avant dans un (très) long rapport.

26 chercheurs provenant de 14 institutions d'horizons variés ont publié la semaine dernière un texte commun d'une centaine de pages dans lequel ils entendent mettre en garde contre une potentielle « *utilisation malveillante de l'intelligence artificielle* ».
Parmi les intervenants, on retrouve des membres de plusieurs universités, notamment celle d'Oxford, de **l'alliance OpenAI** pour la recherche ouverte sur l'IA, du **Centre pour l'étude des risques existentiels** (CSER) et de l'**Electronic Frontier Foundation** (EFF) pour ne citer qu'eux.

Encore un bilan sur les risques de l'intelligence artificielle

Baptisé « *L'utilisation malveillante de l'intelligence artificielle : prévision, prévention et atténuation* », il distille les résultats d'un atelier du Future of Humanity Institute organisé en février de l'année dernière. Le thème principal était justement les risques potentiels de l'intelligence artificielle, un sujet anxiogène pour certains.
On y retrouve également des recherches supplémentaires de la part des auteurs, ainsi que des contributions externes. Ils remercient par exemple des employés de Google/DeepMind, de Microsoft, des chercheurs de l'université de Carnegie-Mellon, des philosophes (Toby Ord et Nick Bostrom), le spécialiste de la nanotechnologie Kim Eric Drexler, etc.

Ce large bilan ne veut pas uniquement alerter des potentiels dangers, mais aussi et surtout ouvrir le débat avec le public, les dirigeants, les chercheurs et les responsables politiques. Ce n'est pas le premier rapport du genre, loin de là.

Pour rappel, au cours des derniers mois, nous nous sommes penchés sur l'**épais rapport de la CNIL**, les **enjeux économiques et cadres légaux soulevés par le Sénat**, le **rapport européen sur les risques de l'intelligence artificielle**, le **bilan #FranceIA** de France Stratégie et du CNNum, la très longue étude de 275 pages de l'OPECST et les **« 23 principes d'Asilomar »**.

Une évaluation des craintes pour les prochaines années

Concernant l'impact – surtout négatif dans le cas présent – de l'intelligence artificielle dans le monde, trois axes principaux sont développés dans ce nouveau rapport : numérique, physique et politique.

Après une présentation générale, les chercheurs nous proposent une projection dans le futur de ce que pourraient être des cas pratiques de dérives liées ou aidées par l'intelligence

artificielle. Des idées de prévention et d'atténuation des risques sont également mises en avant.

Après une lecture attentive de l'ensemble, voici notre bilan détaillé de ce qu'il faut en retenir.

Une étude soutenue financièrement par le Future of Life Institute

Avant d'entrer dans le vif du sujet, prenons un peu de temps pour évoquer la provenance de cette étude publiée par le Future of Humanity Institute de l'université d'Oxford. Près d'un tiers des coauteurs sont d'ailleurs rattachés à cette dernière, quatre autres viennent du **Centre pour l'étude des risques existentiels** et trois de l'association OpenAI.
À elles trois, ces institutions ayant un but assez proche rassemblent donc plus de la moitié des signataires du rapport. Le premier est en effet un institut de recherche interdisciplinaire essayant de répondre aux questions d'ordre général sur l'humanité et son avenir (un vaste sujet), quand le second se penche sur les risques et les menaces liés à la technologie et que le troisième veut une IA bénéfique pour l'humanité.

À la fin du rapport, il est également indiqué que « *ce travail a été soutenu en partie par une subvention du Future of Life Institute* ». Cette association, basée dans la région de Boston, s'est fixée pour mission de « *catalyser et appuyer la recherche et les initiatives visant à sauvegarder la vie et à développer des visions optimistes du futur* ». Stephen Hawking et Elon Musk sont parmi les membres fondateurs, deux personnalités connues pour leurs positions bien tranchées sur les risques liés à l'intelligence artificielle.
En février de l'année dernière, le Future of Life Institute organisait d'ailleurs une grande rencontre baptisée Beneficial AI. Il en était ressorti une liste des « 23 principes d'Asilomar » (voir **notre analyse**) sur les craintes autour du développement de l'intelligence artificielle, signée par plus de 2 500 personnes, dont plus de 1 200 chercheurs. Bref, bon nombre des intervenants ont une vision assez pessimiste du potentiel de l'intelligence artificielle, ce qui se ressent dans le rapport, même s'il essaye de ne pas être trop anxiogène.
Cette étude se penche en effet surtout sur les risques de l'IA, laissant de côté son potentiel et ses avantages. Elle n'en reste pas moins intéressante dans le sens où elle se focalise sur du court terme (les cinq prochaines années au maximum), en essayant de ne pas pousser le bouchon trop loin.

L'intelligence artificielle, une force à double tranchant

Depuis quelques années, l'intelligence artificielle et l'apprentissage automatique reviennent en force sur le devant de la scène, notamment grâce aux puissances de calculs dont on dispose désormais pour un coût réduit et à la quantité de données accessibles pour les entrainer.

Si les futures possibilités de l'IA sont régulièrement mises en avant, parfois sous forme fantasmée, c'est plus rarement le cas des plausibles détournements malveillants des algorithmes, du moins selon les chercheurs. Il ne s'agit pas cette fois d'anticiper l'arrivée d'une IA forte (avec une conscience) souhaitant détruire l'humanité, ou de robots tueurs

comme dans la dernière saison de *Black Mirror*, mais de se concentrer « *sur les types d'attaques que nous sommes susceptibles de voir bientôt arriver si des défenses adéquates ne sont pas développées* ».
Sur le côté obscur du développement de l'IA, les 26 chercheurs anticipent trois axes de développement : une expansion des menaces existantes, l'arrivée de nouveaux risques et une modification du paysage actuel. « *Nous croyons qu'il y a des raisons de s'attendre à ce que les attaques permises par l'utilisation croissante de l'IA soient particulièrement efficaces, ciblées, difficiles à attribuer et susceptibles d'exploiter les vulnérabilités des systèmes IA* » ajoutent-ils.
Problème, toujours selon l'étude, à l'instar d'un maitre Jedi enseignant la force à un apprenti, un chercheur ne peut deviner à l'avance si ses travaux serviront ou non à renforcer le côté obscur de la force (c'est-à-dire des cyberattaques). Quelques exemples de détournements parmi d'autres : des logiciels de détection (et d'exploitation) de failles, les drones autonomes, les bots de communication, etc.

Quand l'IA dépasse les humains...

Dans certaines situations, l'intelligence artificielle dépasse déjà l'humain (**échec et jeux de Go** par exemple), mais pas encore dans toutes, loin de là même. Les chercheurs ajoutent qu'il « *ne semble y avoir aucune raison valable de penser que les performances humaines actuellement observées sont le plus haut niveau possible, même dans les domaines où les performances maximales ont été stables tout au long de l'histoire récente* ». Bref, rien n'est perdu pour l'IA dans tous les domaines où l'Homme occupe pour le moment la première place.
Autre « avantage » des systèmes d'intelligence artificielle : ils « *peuvent augmenter l'anonymat et la distance psychologique* [...] *Par exemple, quelqu'un qui utilise un système d'armes autonome pour effectuer un assassinat, plutôt que d'utiliser une arme de poing, évite à la fois la nécessité d'être présent sur les lieux et le besoin de regarder sa victime*».
Enfin, les travaux de recherche sur l'intelligence artificielle sont souvent librement accessibles à la communauté, facilitant ainsi leur récupération discrète par des personnes malveillantes. Dans le même temps, le coût des systèmes informatiques pour les exploiter baisse grandement au fil des années, permettant donc d'en profiter toujours plus facilement pour des petites organisations ou des particuliers.

... et peut même prendre sa parole

Un autre vecteur d'attaque possible grâce à l'intelligence artificielle est l'usurpation de l'identité vocale. « *Par exemple, la plupart des gens ne sont pas capables d'imiter les voix des autres de façon réaliste ou de créer manuellement des fichiers audio qui ressemblent à des enregistrements de discours humain* ». Nous ne parlons pas ici des sociétés avec d'importants moyens, mais bien d'un internaute lambda ou presque.
Or, les études et les publications se multiplient sur ce point, avec des systèmes (déjà commercialisés) capables d'imiter les humains à partir d'une source audio. Dans la guerre de l'information, ce genre d'outils peut faire des ravages entre de mauvaises mains, permettant de faire dire n'importe quoi à une personne.

On peut également imaginer de faux articles de presse et autres déclarations écrites/vocales reprenant le ton et la présentation des originaux, le tout diffusé sur les réseaux sociaux par exemple.

Numérique, physique, politique : trois domaines à risques dans le futur

L'étude se hasarde à quelques prévisions pratiques. Les chercheurs expliquent qu'ils se basent sur l'état actuel de développement de l'intelligence artificielle, ou sur ce qu'ils estiment possible à court terme (5 ans) afin d'éviter d'entrer dans des suppositions trop hasardeuses.

Ils précisent que certains scénarios se « *produisent déjà sous forme limitée aujourd'hui* », mais qu'ils pourront être largement renforcés à l'avenir. De plus, de nouveaux vecteurs d'attaques auxquels on n'a pas encore pensé seront également mis sur pied. Il ne s'agit donc ici que d'exemples dans les mondes numérique, physique et politique (nous y reviendrons), pas d'une étude sur l'ensemble des risques possibles et imaginables.
Dans le cas du monde numérique, les chercheurs évoquent un phishing très élaboré et parfaitement ciblé en fonction des centres d'intérêt et de la position géographique d'une personne, afin de l'envoyer sur une page dédiée pour la piéger le plus discrètement possible. Une intelligence artificielle peut également traduire dans de nombreuses langues les emails destinés à hameçonner les internautes, facilitant le travail des pirates.

Si le résultat n'est pas encore parfait aujourd'hui (il est souvent très facile d'identifier les tentatives de phishing), il s'améliore grandement au fil des années et cela ne devrait pas se calmer. Est aussi évoqué le cas d'un malware extrêmement virulent cherchant à infecter un maximum de machines à l'aide d'une base de données de vulnérabilités régulièrement et automatiquement mise à jour.

Dans le cas des risques dans le monde physique, l'étude met en avant le cas d'un robot ménager détourné de ses fonctions pour identifier et tenter de tuer une personne. Placé dans le parking souterrain d'un ministère, il attend le passage d'autres robots de la même marque pour se mélanger discrètement à la cohorte.

Une fois fondu dans la masse, il réalise les mêmes tâches que ses petits camarades jusqu'au moment où la « cible » (un ministre dans le cas présent) est identifiée via un logiciel de reconnaissance visuelle, des solutions facilement accessibles sur Internet. Le robot s'approche alors de la personne et explose pour tenter de la tuer.

Des fake news aux algorithmes prédictifs

Dans le monde de la politique, les « fakes news » et autres bots sur les réseaux sociaux pourraient prendre une tout autre ampleur avec l'intelligence artificielle, mais ce n'est pas le seul risque. Voici un autre exemple : une personne lassée d'entendre parler de cyberattaques, de drones et de corruptions dans le monde consulte des articles sur Internet, certains étant des fake news abondant dans son sens. Elle laisse des commentaires sur différents sites pour faire-part de son mécontentement et lance un appel public à la protestation.

Elle passe quelques commandes sur Internet, dont des fumigènes pour terminer avec « panache » son discours. Le lendemain, à son travail, la police arrive : « *notre système prédictif de perturbation civile vous a signalé comme une menace potentielle* ». « *C'est ridicule*, proteste la personne. *Vous ne pouvez pas discuter avec 99,9 % de précision,* surenchérissent les forces de l'ordre*. Maintenant, venez, je ne voudrais pas utiliser la force* ».
Un Précrime sans précogs, mais avec des algorithmes.

Et ce n'est que le début...

Il ne s'agit ici que de quelques exemples anticipant ce que l'intelligence artificielle pourrait devenir si elle continue à se développer de cette manière. Bien évidemment, rien ne dit que cela arrivera sous cette forme, mais c'est une possibilité.

« *Si ces tendances se poursuivent au cours des cinq prochaines années, nous nous attendons à ce que la capacité des attaquants à causer des dommages avec les systèmes numériques et robotiques augmente considérablement* », note le rapport. Il faut également ajouter « *de nouveaux développements, y compris des technologiques sans lien avec l'IA, qui pourraient finalement avoir plus d'impact que les capacités considérées dans ce rapport* ».
Dans tous les cas, ce rapport veut « *sensibiliser le public aux enjeux et à leur importance, et établir un premier programme de recherche* » (nous y reviendrons). Les étapes suivantes nécessiteront « *un engagement de la part de personnes et d'organisations ayant une expertise pertinente. Des ressources monétaires supplémentaires, publiques et privées, contribueraient également à susciter l'intérêt et à attirer l'attention des communautés de recherche concernées* ».
Néanmoins, « *toute prévision fiable à long terme est impossible à faire, car des incertitudes importantes demeurent concernant les progrès des diverses technologies, les stratégies adoptées par les acteurs malveillants et les mesures qui devraient et seront prises par les principales parties prenantes* ».
Rappelons enfin que bon nombre des coauteurs et la Fondation of Life Institute ayant participé au financement de cette étude sont plutôt inquiets vis-à-vis du développement de l'intelligence artificielle. Leur position n'est donc pas surprenante, mais elle reste intéressante à analyser puisqu'elle soulève un débat.