DOC'EN POCHE

PLACE AU DÉBAT

L'intelligence artificielle, avec ou contre nous ?

LE LIVRE BLANC DE L'IA

Rodolphe Gelin
Olivier Guilhem

LE LIVRE NOIR DE L'IA

La documentation Française

Table des matières

Chapitre 4

Liberté et libre arbitre

> *// Si l'intelligence artificielle ne sera sans doute jamais dotée d'une volonté propre et hostile à l'homme, elle est déjà un outil terriblement efficace pour entraver notre liberté de penser et d'agir. Le plus effrayant est que, tel l'explorateur s'enfonçant davantage dans les sables mouvants en voulant s'en extraire, l'utilisateur de l'IA restreint lui-même sa liberté au rythme survolté de ses interactions avec les réseaux sociaux. //*

Les déviances susceptibles de caractériser l'intelligence artificielle sont nombreuses. Nous avons eu l'occasion de nous initier aux écueils inhérents aux principes sur lesquels repose l'IA dans le chapitre 2, traitant des coulisses de cette technologie. Ces risques ne s'arrêtent pas simplement à la phase de conception de l'IA. L'objectif poursuivi par ses créateurs peut être en lui-même pernicieux. Il est quasi impossible de passer en revue toutes les utilisations néfastes de l'outil IA. De la prédiction de l'orientation sexuelle d'un individu au développement de systèmes d'armes létales autonomes, l'esprit humain est hélas bien trop imaginatif quand il s'agit de nuire à son prochain pour espérer parvenir à un recueil complet de ces dangers. L'objectif de ce chapitre est de mettre en exergue certains risques que l'IA peut faire peser sur notre autonomie et notre libre arbitre.

Un rapide regard posé sur ce qui nous entoure nous permet en effet de comprendre la puissante influence exercée par l'IA sur nos choix et sur nos réflexions.

L'information dont nous disposons est traitée par les algorithmes d'une IA à chaque fois que nous utilisons un moteur de recherche ou un réseau social. Nous nous en remettons à notre système de navigation pour prendre le meilleur chemin et c'est aussi une IA qui nous trouve le meilleur billet d'avion, la meilleure location, le meilleur emploi et, même parfois, le meilleur partenaire de vie. Nous vivons dans un monde « algorithmiquement » dépendant, alors que nous ne connaissons généralement pas le fonctionnement de ces algorithmes.

L'IA facilite notre vie quotidienne au point de nous faire adopter à terme un « prêt à penser » dangereux nous transformant en des êtres au sens critique engourdi, au libre arbitre diminué et à l'autonomie restreinte. Cette crainte est justifiée par le développement de certaines pratiques théorisées par exemple sous le principe des « bulles filtrantes », de la théorie du *nudge* (*cf. infra*) ou bien encore du concept de « société de la notation ».

Coincés dans notre bulle

La notion de « bulle de filtres » ou « bulle filtrante » est fondée sur un phénomène psychologique connu, selon lequel l'être humain a tendance à être attiré par ce qui lui ressemble et à entretenir des relations avec des personnes à son image. Cette similitude s'entend aussi bien pour des goûts vestimentaires, musicaux et philosophiques que politiques. Une fois

ce ressort psychologique bien compris, il n'y a plus qu'à développer un algorithme qui définira l'environnement de la navigation internet de l'utilisateur en fonction de ses goûts. Cette personnalisation se construit au gré de ses visites, des « amis » fréquentés sur les réseaux sociaux et autres photographies qui auront été appréciées (« *likée*s »).

Cela peut paraître utile de se voir proposer des films et des musiques correspondant à nos goûts. C'est déjà plus irritant lorsque l'on continue de recevoir des publicités ciblées de chaussures une semaine après en avoir acheté une paire sur un site en ligne. Cela prend encore une autre dimension si l'algorithme s'assure qu'aucun contenu qui puisse nous contrarier ne nous soit transmis. Hors de question, par exemple, de vous présenter un article contraire à vos opinions, de vous mettre en relation avec un groupe de militants pour la préservation des lapins alors que vous êtes chasseur. L'objectif est de « filtrer » votre univers numérique pour vous maintenir dans une bulle de bien-être reflétant ce qui a été « capté » de vous et de votre personnalité. Ce phénomène décrit par le militant et entrepreneur Eli Pariser dans son ouvrage *The Filter Bubble* (2011) a de multiples conséquences. Il vous tient éloigné d'opinions contraires et d'une certaine réalité, vous conforte dans vos opinions tout en polarisant les positions de chacun. Cela a été mis en avant par le juriste et philosophe américain Cass R. Sunstein dès 1999 dans l'article « The Law of Group

Polarization » (University of Chicago Law School). Ce phénomène de polarisation se traduit par l'adoption, par un groupe, de positions plus radicales que celles de ses membres pris individuellement. Il existe donc tout à la fois un mécanisme d'enfermement (on ne peut et ne veut pas sortir de sa bulle), d'entraînement (par le groupe) et d'emballement idéologique (surenchère des membres du groupe).

Un phénomène addictif de « tribalisation » se met en place. La société se scinde de façon manichéenne en clans adverses confortés dans leurs propres certitudes. Nous pourrions avancer que l'individu, tout comme l'IA, y trouve satisfaction. Le premier, en tendant à vouloir être perçu favorablement, trouve un contentement à soutenir son groupe. L'IA remplit sa mission en servant une réalité qui captive l'intérêt de sa cible et en l'enfermant dans un monde binaire bien plus simple à appréhender que le « vrai » monde.

Les IA de « ciblage » ainsi développées ne laissent que peu de possibilités à l'individu de se confronter à une autre réalité, de s'ouvrir à d'autres idées et de favoriser des échanges apaisés. *In fine*, l'algorithme n'est là que pour fournir un service de « vérité » à la demande. Cette information distribuée n'est pas nécessairement frappée du sceau de l'objectivité ou de l'argumentation. Partant des goûts ou des attentes de l'utilisateur, elle peut chercher à l'influencer…

Une société sous influence

Le conte philosophique *Candide ou l'Optimisme*, écrit en 1759 par Voltaire, n'a pas pris une ride. Chassés de cet imaginaire « meilleur des mondes », nous voilà projetés dans un monde digital où notre candeur disparaît au gré de nos usages. Tout semble être fait pour nous guider dans nos choix et nous influencer dans nos décisions. Les « influenceurs » et le développement du *nudge* en sont des exemples parfaits.

Le terme « influenceur » est explicite. Il s'agit par essence, d'une personne exerçant une « influence » auprès des personnes qui suivent ses faits et gestes sur les réseaux sociaux, des *followers,* et qui constituent une communauté cible. Citons par exemple la Française Enjoy Phoenix (de son vrai nom Marie Lopez), qui prodigue des conseils de beauté à près de 3,65 millions d'abonnés sur Youtube et de 5 millions de *followers* sur Instagram (chiffres d'octobre 2020). Bien évidemment, les choses seraient simples si ces influenceurs se présentaient comme tels, à la façon d'un joueur de tennis qui, dans une publicité, nous influence en mangeant une barre chocolatée. Dans les faits, c'est parfois loin d'être le cas. Il peut être difficile de déterminer si certaines de ces « stars » du web et autres réseaux sociaux sont sponsorisées par des marques cherchant à placer leurs produits ou s'ils expriment une véritable opinion indépendante. La frontière de la neutralité est souvent franchie, celui du respect des règles en matière du droit du travail aussi.

Ceci a notamment conduit au dépôt d'une proposition de loi sur « l'exploitation commerciale de l'image d'enfants de moins de seize ans sur les plateformes en ligne », en deuxième lecture début octobre 2020 au Parlement, et visant les mineurs de moins de 16 ans exerçant une activité d'influenceur.

Les organisations professionnelles s'organisent aussi pour tenter d'apporter une certaine forme de déontologie dans ce domaine. La Fédération française des industries Jouet-Puériculture (FJP) et la Fédération des commerces spécialistes des jouets et des produits de l'enfant (FCJPE) ont ainsi établi une charte de bonne conduite des collaborations avec les influenceurs mineurs.

En poussant un peu plus loin la digitalisation et le marketing d'influence, sont apparus des influenceurs virtuels. « Née » en 2007, Hatsune Miku est, avec environ 2,3 millions d'abonnés sur Facebook en octobre 2020, l'une des premières influenceuses virtuelles. Ce personnage commercial, créé à l'origine pour mettre en valeur un logiciel de synthèse vocale, est devenu une véritable *pop star* virtuelle de la chanson qui pourrait s'enorgueillir, si elle était un être humain, de s'être produite au Zénith de Paris en janvier 2020 après avoir collaboré avec Louis Vuitton en 2013. Lil Miquela, apparue en 2016, est un personnage tout aussi virtuel, suivi sur Instagram par plus de 2,7 millions de personnes à la même date. Elle a été créée par la société américaine Brud, et a

Le personnage virtuel Hatsune Miku lors de la répétition d'un spectacle de théâtre kabuki, en 2016 à Tokyo (Japon).

participé à des partenariats avec des marques comme Samsung et Prada. Il existe une multitude d'autres êtres numériques. Ces avatars, dont les créateurs et les intentions ne sont pas toujours identifiés, sont de véritables agents, outils d'influence.

Des agences marketing se sont spécialisées dans le *coaching* de ces vedettes numériques. À titre d'exemple, la société ID-agencesdesmediassociaux.com offre, au travers de son site « agence des

influenceurs », un service dédié au marketing d'influence. D'autres se dédient au montage d'outils et de campagnes d'influence usant, après avoir ciblé l'influenceur et le message adéquats, du pouvoir de ce dernier auprès de son auditoire. Ces agences sont en fait les influenceurs de l'ombre.

Les fruits de recherches menées dans de multiples disciplines ont mis à profit l'IA pour augmenter l'efficacité des instruments d'influence. L'économie comportementale, qui étudie les comportements des individus dans leur prise de décisions économiques, en est une illustration. Est notamment issu de ces travaux le concept de *nudge* (« coup de coude », mais traduit par « coup de pouce » en français). Derrière ce terme se cachent l'ensemble des techniques qui consistent à utiliser les biais cognitifs des individus afin de les inciter à changer de comportement sans les contraindre. Ces techniques furent mises en lumière par Richard Thaler, prix Nobel d'économie en 2017 pour sa « compréhension de la psychologie de l'économie » et Cass Sunstein avec la publication, en 2008, de leur ouvrage *Nudge ? Émotions, habitudes, comportements : comment inspirer les bonnes décisions*.

Ces modes d'incitation douce, de « *design* » comportemental, vont par exemple nous encourager à rester connectés plus longtemps à une application (afin d'obtenir plus de points, etc.) ou à rendre un bien loué en parfait état (pour éviter une mauvaise note). Si l'IA est capable d'envoyer un message ciblé,

le *nudge* peut, quant à lui, adapter ce message à sa cible pour le pousser subtilement à l'action. La théorie du *nudge* vise à lever les barrières psychologiques. Les exemples bénéfiques du *nudge* sont généralement cités (incitations à prendre les escaliers plutôt que les ascenseurs, à placer une « cible » dans les urinoirs pour améliorer la précision de l'utilisateur, etc.) ; mais ces modes de « communication » sont pour l'instant assez opaques et pas réellement évalués et encadrés.

Certaines sociétés combinent ainsi « marketing cognitif » et IA afin de développer de véritables outils de persuasion et de propagande. C'est le cas notamment de la très célèbre société britannique Cambridge Analytica, active en 2016 lors du référendum sur le Brexit *via* l'entreprise canadienne AggregateIQ, mais aussi lors de l'élection présidentielle américaine de la même année. Les outils de ciblage et de manipulation mis en œuvre influèrent sur le vote de nombre d'électeurs. Les activités de cette société auraient touché 68 pays. Si Cambridge Analytica n'existe plus aujourd'hui, de nombreuses autres officines demeurent et continuent d'exercer cette activité d'influence. L'organisation russe The Internet Research Agency (IRA) a, par exemple, été accusée d'avoir créé sur les réseaux sociaux des centaines de faux profils en vue d'exacerber les tensions sociales aux États-Unis lors des scrutins présidentiels. Comme le rapporte le *Washington Post* dans un article du 16 février 2018, l'IRA (mais pas seulement elle) est

mise en accusation par le Département de la justice américain pour son interférence lors de la campagne présidentielle de 2016. Cette même organisation semble avoir repris du service pour la campagne américaine de novembre 2020. Selon une information de *France 24* du 2 septembre 2020, le FBI, Twitter et Facebook ont ainsi mis fin à une nouvelle opération de propagande montée par l'IRA.

Une étude menée par l'Institut internet de l'université d'Oxford publiée en septembre 2019 recense le nombre de pays frappés par des campagnes de désinformation. De 28 en 2017, ce chiffre passe à 48 en 2018 pour atteindre 70 en 2019. Il est aisé d'arguer que la réclame publicitaire, le *lobbying* et autres techniques de manipulation politique ont toujours existé. Ceci ne saurait en rien justifier le développement de pratiques que l'on peut considérer *a minima* comme non éthiques.

Michal Kosinski, professeur associé en comportement organisationnel à l'université de Stanford, est à l'origine de travaux portant sur les traces numériques que nous laissons au quotidien sur la Toile. Ses études ont mis en évidence que ces données étaient des révélateurs de notre personnalité, de nos activités et de nos désirs, permettant de prévoir certains de nos futurs comportements. Comme le relève M. Kosinski, ceci peut être utilisé à bon ou à mauvais escient, notamment pour manipuler les individus. Ses travaux ont inspiré des sociétés comme Cambridge Analytica. Le

chercheur alerte sur les dangers d'un algorithme qui pourrait révéler nos traits les plus intimes. Il considère que, dans un proche avenir, il sera non seulement possible d'influencer l'opinion politique d'une personne mais aussi ses émotions.

L'IA permet, par ses capacités, le développement d'opérations d'influence de plus en plus ciblées et pertinentes. Ces dernières peuvent s'effectuer, grâce à elle, à moindre coût et en s'affranchissant progressivement des barrières linguistiques et culturelles qui constituaient auparavant des sortes de protections naturelles. Il n'y aurait à terme aucune échappatoire aux algorithmes. Nous serions réduits à n'être qu'une variable influençable.

Cette situation est en partie rendue possible par le fait que seul un petit nombre d'acteurs délivre une part importante de l'information à laquelle nous accédons. Les GAFAM (Google, Amazon, Facebook, Apple et Microsoft) peuvent être considérés comme des sociétés technologiques mais aussi comme des médias. Fin 2019, le média social Facebook comptait environ 2,5 milliards d'utilisateurs actifs. Il délivrait donc approximativement à un tiers de l'humanité une information choisie et ciblée par ses algorithmes. Cela pose des questions de souveraineté au regard de l'hégémonie de ces géants technologiques gérant une très grande partie de nos communications et de l'information mondiale. En plus de contrôler ces flux, ces sociétés les commercialisent. Détentrices

de leurs propres canaux de diffusion, elles opèrent aussi comme de véritables régies publicitaires. Elles commercialisent les données de leurs utilisateurs auprès d'annonceurs, qui payent pour leur envoyer un message calibré. Message qui, en retour, engendrera un surcroît de données qui pourront de nouveau être monétisées.

La surveillance croissante exercée par les IA de ciblage absorbant un nombre grandissant de données toujours plus précises (et donc valorisables) ainsi que la délivrance de messages ajustés pour satisfaire annonceurs et utilisateurs renforcent sans cesse le cycle de l'influence. Ce modèle incite à s'interroger sur la déontologie, l'objectif poursuivi ainsi que sur l'objectivité des informations que l'IA sélectionne pour nous.

L'essor des *deepfakes* (ou hypertrucages) devrait encore accentuer le phénomène. Le *deepfake* utilise la technique des GAN (voir chapitre 2) pour, par exemple, remplacer le visage et la voix d'une personne par un autre visage, une autre voix. Cette technique peut ainsi permettre de cloner une voix, d'altérer une vidéo ou une image. Les personnes mal intentionnées pourraient donc faire tenir des propos fantaisistes à des hommes politiques et autres capitaines d'industrie. Un navigateur de recherche donnera aisément accès à de fausses vidéos de Barack Obama ou de Donald Trump, mais aussi d'Emmanuel Macron ou de Nicolas Cage, leur faire proférer des propos qu'ils n'ont jamais tenus. En juillet 2018, le sénateur démocrate Mark

Warner formulait dans un *livre blanc* des propositions pour lutter contre la désinformation sur les réseaux sociaux, et notamment les *deepfakes*. L'alerte est aussi donnée en France au travers notamment d'un rapport intitulé *Les manipulations de l'information, un défi pour nos démocraties*.

Publié en septembre 2018 par le Centre d'analyse, de prévision et de stratégie du ministère des Affaires étrangères et l'Institut de recherche stratégique de l'École militaire (IRSEM), ce rapport prend en considération l'IA en tant que facteur de risque mais aussi comme une opportunité permettant de lutter contre la désinformation.

Cette technique du *deepfake* ouvre la voie à d'autres dangers potentiels, comme le contournement de systèmes de sécurité ou l'usurpation d'identité. En mars 2019, la filiale anglaise d'un groupe allemand du secteur de l'énergie a été victime d'une « fraude au président ». Après avoir cloné la voix du président de la maison-mère, les escrocs ont réussi à persuader cette société britannique d'effectuer un versement frauduleux de près de 220 000 euros.

Une société de la notation

Le meilleur moyen de collecter des données afin d'améliorer continuellement les services rendus par l'IA est de mettre en place un modèle où l'utilisateur serait contributeur.