Algor-éthique : intelligence artificielle et réflexion éthique

Paolo Benanti, traduit de l'italien par **Alain Thomasset**, avec l'aide de l' **IA DeepL**

Dans Revue d'éthique et de théologie morale 2020/3 (N° 307), pages 93 à 110
Éditions Éditions du Cerf

ISSN 1266-0078
ISBN 9782204138086
DOI 10.3917/retm.310.0093

Paolo Benanti

Algor-éthique : intelligence artificielle et réflexion éthique

De l'éthique de la technologie à l'algor-éthique

L'intelligence artificielle (IA) est en train de changer le monde : chaque activité humaine, de la médecine à la sécurité nationale, est en train de subir de profondes mutations. Les systèmes équipés d'IA n'aident pas seulement les humains, mais dans de plus en plus de situations, ils donnent naissance à des systèmes, des robots ou des bots complètement autonomes. Face à ce flot d'intelligence artificielle, la question éthique est urgente. Plus l'IA devient universelle, plus il est nécessaire de développer un nouveau langage universel qui puisse gérer l'innovation[1].

Nous devons commencer par dissiper certains malentendus éventuels. L'un des malentendus les plus courants lorsqu'on parle d'éthique est qu'on pense que l'éthique est une sorte de chaîne qui devrait fixer des limites à la liberté. Parler d'éthique de la technologie reviendrait alors à essayer de tracer des limites a priori à la technologie. Mais est-ce bien le cas ? Afin de comprendre ce que signifie l'éthique de la technologie, nous devons retracer un parcours qui a commencé il y a longtemps.

1. Ce texte retravaille et résume certaines des idées contenues dans les textes suivants : Paolo Benanti, *Digital Age. Teoria del cambio d'epoca. Persona, famiglia e società*, San Paolo 2020 ; Paolo Benanti, *Realtà sintetica. Dall'aspirina alla vita: come ricreare il mondo?*, Castelvecchi, Rome, 2018 ; Paolo Benanti, *Le macchine sapienti*, Marietti, Bologne, 2018 ; Paolo Benanti, *Oracoli. Tra algoretica e algocrazia*, Luca Sossela Editore, Rome, 2018.

L'éthique de la technologie

Selon les anthropologues, il y a 70 000 ans, notre espèce, l'*Homo sapiens*, a quitté l'Afrique australe, le berceau de notre existence, pour coloniser le monde entier. Nous avons atteint chaque endroit d'une manière vraiment unique, en montrant ce qui fait notre spécificité en tant qu'espèce. Jusqu'alors, chaque espèce biologique vivait dans un climat particulier, celui de son habitat. Si un mammouth des steppes sibériennes s'est déplacé en Afrique et en Inde, c'est parce qu'un membre de sa progéniture a subi une mutation génétique, perdant son long pelage et devenant capable de survivre dans les climats chauds du sud. Lorsque l'homme d'Afrique australe s'est déplacé à travers le monde, y compris dans la steppe sibérienne, il n'a pas attendu la naissance d'un descendant au pelage robuste, contrairement au mammouth. En d'autres termes, il n'a pas attendu la naissance d'un *Homo sapiens hipster*. L'homme s'est habillé de la fourrure du mammouth. En d'autres termes, ce qui chez toutes les autres espèces est fourni par le code génétique est modifié pour ce qui nous concerne par l'artefact technologique. Les capacités que possèdent les autres animaux leur sont données par des aptitudes génétiques et ne peuvent changer que si leur ADN change. Nous ne faisons pas cela. Nous coopérons les uns avec les autres, transmettons des informations sur le monde et éduquons les générations suivantes à faire certaines choses grâce à des artefacts technologiques. Si le dauphin peut nager grâce à son ADN, l'homme est différent. L'homme se transforme et transforme le monde dans lequel il vit grâce à des artefacts technologiques. La technologie est le lieu où nous trouvons tout ce condensé. Avec la technologie, nous changeons le monde et nous nous changeons nous-mêmes pour habiter le monde. L'éthique de la technologie n'est rien d'autre que la constitution naturelle de l'artefact technologique.

Les intelligences artificielles sont des artefacts technologiques. Mais différentes de tous les artefacts produits à ce jour. Tous les outils que nous avons produits permettent à l'homme d'effectuer certaines tâches. Des massues primitives aux grandes machines industrielles, tous ces outils ont été utilisés pour effectuer des tâches précises plus rapidement et plus efficacement. L'IA, tant chez les robots que chez les bots, a dépassé le concept d'artefact et de machine que nous connaissions jusqu'à présent. Tous les mécanismes automatiques

que nous avons construits pendant la révolution industrielle l'ont été en tenant compte de leur objectif. Ils ont simplement fait ce pour quoi ils ont été conçus. Aujourd'hui, l'IA n'est pas conçue de cette manière. Ce ne sont pas des logiciels programmés, mais des systèmes entraînés. Elle surpasse le modèle classique « si-ceci-alors-cela » dans lequel un ingénieur en informatique a d'abord prédit toutes les occurrences possibles. L'IA répond de manière autonome à un problème qui lui est posé. Ces artefacts sont une nouvelle espèce de machines. *Machina sapiens*. Aujourd'hui, le monde n'est plus seulement habité par l'*Homo sapiens*, mais aussi par des machines *sapiens*. Si la machine est autonome, qui est responsable de ses décisions ? Qui l'a conçu ? Qui l'utilise ? Qui la vend ? Qui l'a acheté ? Actuellement, une intelligence artificielle moyenne est capable de faire un meilleur diagnostic médical que le médecin moyen. Sommes-nous prêts à déléguer tout ce pouvoir de décision à des machines ? Pour pouvoir répondre à cette question, nous devons clarifier une question fondamentale : une intelligence artificielle peut-elle faire un choix parfait ?

Une intelligence artificielle peut-elle faire un choix parfait ?

Les spécialistes des données nous disent que le problème réside dans la qualité et la quantité des données. Lorsque nous disposons d'une base de données parfaite pour nos services d'analyse d'impact, la machine fait des choix parfaits. Mais est-ce bien le cas ? Nous avons déjà eu cette impression. Laplace a affirmé que si nous connaissions en un instant l'emplacement de toutes les particules que contient l'univers, nous serions en mesure de prédire tout l'avenir, c'est-à-dire de connaître tout le passé de l'univers. C'était le célèbre démon de Laplace. Aujourd'hui, la question est appliquée à l'intelligence artificielle et à ses choix basés sur des données. Quelles sont les données sur lesquelles l'intelligence artificielle s'appuie ? En bref, nous pouvons dire que les données ne sont rien d'autre qu'une carte du monde. Tout ce qui existe dans le monde est cartographié, enregistré et mis dans une base de données avec une carte. Mais une carte peut-elle être une copie exacte du monde ? Oublions la question philosophique sur cette possibilité et abordons-la d'un point de vue opérationnel. Si nous étions capables de créer une carte qui soit la copie exacte de la réalité, y compris tout ce qui s'y trouve, y compris les passants, les

feuilles d'arbres, etc., ce serait en fait aussi complexe que la réalité, trop complexe pour prendre des décisions et donc inutile.

Autrement dit, nous serions confrontés au paradoxe bien connu raconté par Jorge Luis Borges dans un fragment de « La rigueur de la science », le dernier de l'*Histoire universelle de l'infamie* publié pour la première fois en 1935. Comme c'est son habitude, l'auteur argentin attribue la citation à un livre qui n'existe pas réellement :

> En cet empire, l'Art de la Cartographie fut poussé à une telle Perfection que la Carte d'une seule Province occupait toute une ville et la Carte de l'Empire toute une Province. Avec le temps, ces Cartes Démesurées cessèrent de donner satisfaction et les Collèges de Cartographes levèrent une Carte de l´Empire, qui avait le Format de l´Empire et qui coïncidait avec lui, point par point. Moins passionnées pour l´Étude de la Cartographie, les Générations Suivantes réfléchirent que cette Carte Dilatée était inutile et, non sans impiété, elles l'abandonnèrent à l'Inclémence du Soleil et des Hivers. Dans les Déserts de l'Ouest, subsistent des Ruines très abîmées de la Carte. Des Animaux et des Mendiants les habitent. Dans tout le Pays, il n'y a plus d'autre trace des Disciplines Géographiques. (Suarez Miranda, *Viajes de Varones Prudentes*, Livre IV, Chapitre XIV, Lérida, 1658)[2].

Les données sont une carte de la réalité, elles représentent une réduction de la réalité et pour cette raison, elles sont utiles pour prendre des décisions. En outre, l'IA travaille sur des bases de données et des capteurs. Mais même les capteurs ne lisent pas toute la réalité : ils n'en prennent qu'une partie et la transforment en données. Nous sommes ici au point clé de la question. Comme les intelligences artificielles fondent leurs décisions sur des données et qu'elles ne constituent pas une copie parfaite de la réalité, il n'est pas concevable a priori qu'une machine équipée d'une intelligence artificielle puisse faire un choix sans erreur. La machine *sapiens* sera toujours et constitutivement faillible. L'IA a besoin d'une éthique constitutive. Comme l'intelligence artificielle peut faire des erreurs, il est nécessaire de comprendre comment gérer cette erreur. La question de l'éthique est primordiale et urgente. Il faut trouver un système éthique commun afin que l'utilisation de ces systèmes ne produise pas d'injustices, ne nuise pas aux personnes et ne crée pas de forts déséquilibres mondiaux.

2. Jorge Luis Borges, *Histoire universelle de l'infamie*, Pocket, 2003 (original en 1935).

Les lignes directrices d'une algor-éthique

Quelles sont les lignes directrices éthiques qui peuvent nous guider dans la réalisation de ce nouveau langage humain destiné à mettre l'*Homo sapiens* en contact avec la *machina sapiens* ? L'histoire de l'éthique nous aide dans cette recherche.

La première ligne directrice est ce que l'on pourrait appeler la peur de l'incertitude. Quel que soit le choix que nous faisons, nous savons qu'il aura des conséquences. Nous pouvons tous choisir librement, mais ce qui se passe une fois que nous avons fait un choix ne dépend pas toujours de nous. Tout choix libre et conscient comporte un horizon d'incertitude. L'un des principaux paradigmes éthiques est la gestion de l'incertitude. C'est le premier moteur éthique : être conscient que les choix faits peuvent aussi produire des effets indésirables et gérer ce risque.

Une deuxième ligne directrice, très importante à considérer, est la tension entre l'égalité et la recherche du bonheur. Toutes les guerres les plus sanglantes que nous avons connues entre 1800 et 1900 ont été menées pour l'égalité de tous les hommes. En fait, l'utilisation de ces technologies risque de produire de nouvelles inégalités. L'éthique de l'IA doit nous en protéger. La dignité humaine est la valeur éthique, et non pas la valeur des données. De plus, un État acquiert sa légitimité s'il permet à l'individu de mener sa propre recherche du bonheur. Ces nouvelles technologies avec leur possibilité de profilage, avec leur possibilité de prédire le comportement des êtres humains, peuvent en effet rendre très difficile la situation d'une existence individuelle libre. Il ne faut pas seulement considérer le bien et le mal qui peuvent survenir pour l'individu (peur de l'incertain) mais pour la société dans son ensemble : faut-il protéger l'égalité des individus et la possibilité pour chacun de pouvoir poursuivre son propre bonheur ?

Enfin, nous devons être conscients d'une vérité fondamentale. L'éthique seule est fragile. Tout comme la dignité humaine a été piétinée par les régimes totalitaires du XX^e^ siècle parce qu'elle n'était protégée par aucun droit, l'éthique de l'IA risque d'être inefficace si elle ne devient pas une politique contraignante qui protège l'individu et la coexistence sociale.

L'existence des machines *sapiens* appelle un nouveau langage universel qui puisse traduire ces directives éthiques en directives

exécutables par la machine. Mais comment faire ? À l'ère numérique, le monde est régulé par des algorithmes. Plus d'un parle d'une algo-folie. Afin d'éviter que ce domaine de l'algorithme n'existe également grâce à l'IA, nous devons commencer à développer ce langage commun d'« algor-éthique ».

Pour pouvoir développer une algor-éthique, nous devons préciser dans quel sens nous parlons de valeur. En fait, les algorithmes fonctionnent sur des valeurs de nature numérique. L'éthique parle plutôt de valeur morale. Nous devons établir un langage qui puisse traduire la valeur morale en quelque chose de calculable pour la machine. La perception de la valeur éthique est une capacité purement humaine. La capacité à travailler sur des valeurs numériques est plutôt la capacité de la machine. L'algor-éthique naît si nous sommes capables de transformer la valeur morale en quelque chose de calculable.

Mais dans la relation entre l'homme et la machine, le véritable connaisseur et porteur de valeurs est la partie humaine. La dignité humaine et les droits de l'homme nous disent que c'est l'homme qui doit être protégé dans la relation entre l'homme et la machine. Cette évidence nous donne l'impératif éthique fondamental pour la machine *sapiens* : doutez de vous-même. Nous devons permettre à la machine d'avoir un certain sens de l'incertitude. Chaque fois que la machine ne sait pas avec certitude si elle protège la valeur humaine, elle doit exiger une action humaine. Cette directive fondamentale est réalisée par l'introduction de paradigmes statistiques au sein de l'IA. Des tentatives de ce type sont menées par Google et Uber avec des bibliothèques statistiques spécialisées. C'est cette capacité d'incertitude qui doit être au cœur de la décision de la machine. Si la machine, à chaque fois qu'elle est dans une condition d'incertitude, demande à l'humain, alors ce que nous créons est une intelligence artificielle qui place l'humain au centre ou, comme on dit chez les techniciens, une conception centrée sur l'humain. La norme fondamentale est celle qui construit l'IA d'une manière centrée sur l'homme.

À partir de cette grammaire de base, nous pouvons développer un nouveau langage universel : l'algor-éthique. Celui-ci aura sa propre syntaxe et développera sa propre littérature. Ce n'est ni le lieu ni le moment de dire tout ce qui peut être exprimé avec ce langage, mais il nous semble que nous devons au moins donner quelques exemples qui en révèlent le potentiel.

Anticipation – Lorsque deux humains travaillent ensemble, l'un est capable d'anticiper et de soutenir les actions de l'autre en ayant l'intuition de ses intentions. Cette compétence est à la base de la ductilité qui caractérise notre espèce : depuis l'Antiquité, elle permet à l'homme de s'organiser. Dans un environnement mixte, l'IA doit également être capable d'intuition de ce que les hommes veulent faire, et elle doit suivre leurs intentions en coopérant : la machine doit s'adapter à l'homme, et non l'inverse.

Transparence – Les robots travaillent généralement selon des algorithmes d'optimisation : la consommation d'énergie de leurs servomoteurs, les trajectoires cinématiques et les vitesses de fonctionnement sont calculées pour être aussi efficaces que possible dans la réalisation de leur objectif. Pour que l'homme puisse vivre avec la machine, l'action de celle-ci doit être intelligible. L'objectif principal du robot ne doit pas être d'optimiser ses actions, mais de rendre ses actions compréhensibles et intuitives pour l'homme.

Personnalisation – Un robot, grâce à l'IA, est en relation avec l'environnement en ajustant son comportement. Là où l'homme et la machine vivent ensemble, le robot doit également être capable de s'adapter à la personnalité de l'homme avec lequel il coopère. L'*Homo sapiens* est un être émotionnel ; la machine *sapiens* doit reconnaître et respecter cette caractéristique unique et particulière de son partenaire de travail.

Adéquation – Les algorithmes d'un robot déterminent ses lignes de conduite. Dans un environnement partagé, le robot doit être capable d'adapter ses objectifs en observant la personne et donc de comprendre quel est l'objectif pertinent dans chaque situation spécifique. La machine doit, en d'autres termes, acquérir une « humilité artificielle » pour attribuer une priorité opérationnelle aux personnes présentes, et non à la réalisation d'un objectif prédéterminé.

À l'ère de l'IA, ces quatre paramètres sont un exemple de la manière de protéger la dignité de la personne. Le problème est avant tout philosophique et épistémologique. L'IA « fonctionne » selon des schémas qui relient les données. De quel type de connaissance s'agit-il ? Quelle est sa valeur ? Comment doit-elle être traitée et prise en compte ?

En bref, la question qui se pose devant la technologie est éthique et philosophique : dans la mesure où nous voulons confier les compétences humaines, la compréhension, le jugement et l'autonomie

d'action aux systèmes logiciels d'IA, nous devons comprendre la valeur, en termes de connaissances et de capacité à agir de ces systèmes qui se prétendent intelligents et cognitifs.

Aujourd'hui, l'IA se développe soit grâce au marché, soit grâce à l'État. Nous devons réfléchir à d'autres modes. Par exemple, en développant des algorithmes de vérification indépendants qui peuvent certifier ces quatre capacités des machines. Il est également possible d'émettre l'hypothèse que des entités tierces indépendantes, par l'écriture d'algorithmes dédiés, soient capables d'évaluer l'aptitude de l'IA à vivre avec des humains. Ce n'est qu'en respectant ces indications que l'innovation peut être guidée vers un authentique développement humain[3].

Quelle utilisation pour les données de synthèse

Les grandes collections de données visuelles, composées d'images et de vidéos, qui sont l'héritage accumulé des entreprises technologiques les plus puissantes sur le marché de l'IA, constituent un énorme avantage concurrentiel. Ces bases de données ont creusé un fossé qui maintient les progrès de l'apprentissage machine hors de portée de nombreux acteurs. Cet avantage semble destiné à être annulé par l'avènement des données de synthèse. Quels défis cela pose-t-il pour une philosophie et une éthique de l'IA ?

Données visuelles et données de synthèse

Les plus grandes entreprises technologiques du monde, telles que Google, Facebook, Amazon, pour ne citer que quelques-unes des plus importantes, développent la vision par ordinateur et l'IA pour former leurs ordinateurs. Ils collectent d'immenses ensembles de données visuelles composés d'images, de vidéos et d'autres données visuelles auprès de leurs consommateurs et les utilisent pour entraîner leurs algorithmes. Ces ensembles de données constituent un avantage concurrentiel pour les grandes entreprises technologiques : c'est grâce

3. Voir le blog « paolobenanti.com », le 15.02.2019.

à cet atout que les progrès en matière d'apprentissage machine et les processus permettant aux ordinateurs et aux algorithmes d'apprendre plus rapidement restent hors de portée de nombreux concurrents.

Mais si l'on regarde ce qui se passe, il semble que cet avantage pourrait disparaître en raison de la capacité de quiconque à créer et à exploiter des données de synthèse pour former des ordinateurs. Vous pouvez former efficacement des algorithmes avec des données de synthèse dans de nombreux scénarios d'utilisation, notamment dans le commerce de détail, la robotique, les véhicules autonomes, le commerce, etc.

Les données de synthèse sont des données générées par ordinateur qui reproduisent – ou peut-être serait-il préférable de dire simulent – des données réelles ; en d'autres termes, il s'agit de données créées par une simulation informatique, et non par un être humain ou une activité réelle. Aujourd'hui, nous sommes capables de concevoir des algorithmes logiciels pour créer des données simulées ou, selon cette expression, des données de synthèse réalistes.

Les scientifiques et les ingénieurs logiciels utilisent des données de synthèse pour apprendre à un ordinateur comment réagir à certaines situations ou à certains critères, en substituant ces données à celles acquises dans des activités du monde réel. Dans le processus de formation, l'un des aspects les plus importants, tant pour les données réelles que pour les données de synthèse, est de disposer d'étiquettes précises afin que les ordinateurs puissent traduire les données visuelles et leur donner un sens.

De nombreuses entreprises utilisent la vision artificielle, l'apprentissage machine et l'intelligence artificielle pour analyser des données visuelles dans tous les secteurs d'activité : santé, robotique, logistique, cartographie, transport, fabrication, etc. De nombreuses start-up, aux idées vraiment innovantes, ont le problème du démarrage à froid (*cold bootstrap*) car elles ne disposent pas assez de données étiquetées de qualité suffisante pour entraîner leurs algorithmes : un système ne peut tirer aucune conclusion concernant les utilisateurs ou les articles sur lesquels il n'a pas encore recueilli suffisamment d'informations. Les entreprises en démarrage peuvent collecter des données pertinentes ou en même temps collaborer avec d'autres pour rassembler des données pertinentes. Par exemple, les détaillants peuvent être contactés pour des données sur le comportement

d'achat ou les hôpitaux pour des données médicales. De nombreuses start-up, lorsqu'elles en sont à leurs débuts, essaient de résoudre leur problème de démarrage à froid en créant des simulateurs de données pour générer des données contextuellement pertinentes avec des étiquettes de qualité pour entraîner leurs algorithmes.

Les grandes entreprises technologiques ne rencontrent pas ce problème car elles ont dans leur activité propre des sources exponentielles pour collecter des données uniques et contextuellement pertinentes.

Des avancées rapides

Les progrès réalisés avec les données de synthèse sont remarquables. Serge Belongie, professeur à Cornell Tech, qui s'occupe de vision par ordinateur depuis 25 ans, interrogé sur le sujet, a déclaré :

> Dans le passé, notre champ de vision informatique a jeté un regard méfiant sur l'utilisation des données synthétiques, car elles étaient visiblement trop fausses. Malgré les avantages évidents d'obtenir gratuitement des informations véridiques par la vision, notre inquiétude était de former un système qui fonctionnerait très bien en simulation mais qui échouerait lamentablement dans la nature. Aujourd'hui, le jeu a changé : l'écart entre la simulation et la réalité disparaît rapidement. Au minimum, nous pouvons préformer des réseaux neuronaux convolutionnels très profonds sur des images quasi photoréalistes et les affiner sur des images réelles soigneusement sélectionnées[4].

Par exemple, AiFi est une start-up, en phase initiale de développement, qui construit une plateforme de vision artificielle et d'intelligence artificielle pour fournir une solution de caisse plus efficace et sans personnel tant pour les magasins familiaux que pour les grandes chaînes. Ils mettent en place une solution de paiement sans enregistrement, similaire à Amazon Go. La solution de création de données de synthèse d'AiFi est également devenue l'un de leurs avantages technologiques spécifiques. Grâce au système AiFi, les acheteurs pourront entrer dans un magasin de détail et récupérer des articles sans avoir à

4. Voir Evan Nisselson, "Deep learning with synthetic data will democratize industry", https://techcrunch.com/2018/05/11/deep-learning-with-synthetic-data-will-democratize-the-tech-industry/.

utiliser de l'argent liquide, une carte ou à scanner des codes-barres. Ces systèmes intelligents devront surveiller en permanence des centaines ou des milliers d'acheteurs dans un magasin et les reconnaître ou les « ré-identifier » au cours d'une séance d'achat complète.

Ying Zheng, cofondatrice et directrice scientifique d'AiFi, avait auparavant travaillé avec Apple et Google. La femme d'affaires explique :

> Le monde est vaste et peut difficilement être décrit par un petit échantillon d'images et de signaux réels. Sans parler du fait que l'acquisition de signaux de haute qualité prend du temps et de l'argent, et que c'est parfois impossible. Avec des données de synthèse, nous pouvons saisir un aspect du monde, petit mais pertinent, avec des détails parfaits. Dans notre cas, nous créons des simulations de magasin à grande échelle et nous produisons des images de haute qualité avec des balises parfaites par pixel et nous les utilisons pour former avec succès nos modèles d'apprentissage profond. Cela permet à AiFi de créer à grande échelle des solutions sans caisse et supérieures[5].

La robotique est un autre domaine qui utilise des données de synthèse pour former les robots à diverses activités dans les usines, les entrepôts et dans toute l'entreprise. Josh Tobin est chercheur à OpenAI, une société de recherche en intelligence artificielle à but non lucratif qui vise à promouvoir et à développer l'intelligence artificielle au profit de l'humanité tout entière. Tobin fait partie d'une équipe qui travaille sur la construction de robots d'apprentissage. Ces derniers se sont entraînés entièrement avec des données simulées et les résultats ont ensuite été transférés sur un robot physique, qui peut maintenant, de manière presque incroyable, apprendre une nouvelle tâche après avoir vu une action effectuée une seule fois. Ils ont développé et mis en œuvre un nouvel algorithme appelé « apprentissage par imitation en une fois », qui permet aux humains de communiquer la manière d'effectuer une nouvelle tâche en l'exécutant en réalité virtuelle. En une seule démonstration, le robot est capable de résoudre la même tâche à partir d'un point de départ arbitraire, puis de poursuivre la tâche.

L'objectif est d'apprendre les comportements en simulation et de transférer cet apprentissage dans le monde réel. L'hypothèse était de

5. *Ibid.*

voir si un robot peut faire des tâches précises aussi bien à partir des données simulées. Les concepteurs ont commencé par des données simulées à 100 % et ont pensé que cela ne fonctionnerait pas aussi bien que d'utiliser des données réelles pour former les ordinateurs. Cependant, les données simulées pour les tâches robotiques d'entraînement ont fonctionné bien mieux que prévu.

De nombreuses grandes entreprises technologiques, des constructeurs automobiles et des start-up s'efforcent de développer des véhicules autonomes. Les développeurs ont réalisé qu'il n'y a pas assez d'heures dans une journée pour collecter suffisamment de données réelles pour couvrir les kilomètres nécessaires pour apprendre aux voitures à conduire. May Mobility est une start-up qui construit un service de microtransit avec des véhicules autoguidés. Leur P.-D.G. et fondateur, Edwin Olson, parle des données de synthèse :

> L'une de nos utilisations des données de synthèse est l'évaluation des performances et de la sécurité de nos systèmes. Cependant, nous ne pensons pas qu'un nombre raisonnable de tests (réels ou simulés) soit suffisant pour démontrer la sécurité d'un véhicule à conduite autonome. La sécurité fonctionnelle joue un rôle important. La flexibilité et la polyvalence de la simulation rendent la formation et l'essai de véhicules autonomes dans des conditions très variables particulièrement précieux et beaucoup plus sûrs. Les données simulées peuvent également être étiquetées plus facilement car elles sont créées par des ordinateurs, ce qui permet de gagner beaucoup de temps[6].

Jusqu'à présent, les principales entreprises de plateformes numériques ont exploité leurs gisements de données pour maintenir leur avantage concurrentiel. Les données de synthèse sont un élément perturbateur majeur de ces avantages, car elles réduisent considérablement les coûts et la vitesse de développement, ce qui permet à de petites équipes agiles de rivaliser et de gagner.

Le défi et l'opportunité pour les start-up de concurrencer les grandes entreprises sont d'exploiter de meilleures données visuelles avec des étiquettes correctes pour former les ordinateurs avec précision pour différents cas d'utilisation. La simulation de données permettra de réduire la distance qui sépare les grandes entreprises technologiques des

6. *Ibid.*

jeunes pousses. Avec le temps, les grandes entreprises aussi créeront probablement des données de synthèse pour augmenter leurs données réelles, et un jour cela pourrait à nouveau déséquilibrer le terrain de jeu.

Questions philosophiques et éthiques

À ce stade, cependant, la question devient de nature philosophique et éthique. Le premier point est de nature épistémologique. L'IA fonctionne en trouvant une signification et en attribuant des corrélations à de grands ensembles de données. Or, cette signification et sa valeur épistémologique sont en elles-mêmes problématiques et doivent être pleinement comprises. La quantité de connaissances offertes par l'IA et le type de connaissances acquises ont été traités plusieurs fois sur mon blog et sont au centre de mon livre *Oracoli* de la série Collassi de l'éditeur Luca Sossela. Aujourd'hui la question devient plus complexe : la virtualité offre des connaissances sur la réalité avec une revendication de vérité et des orientations sur les actions autonomes des algorithmes d'IA.

Outre ce nouveau point épistémologique, une question éthique plus importante se pose. Si l'IA et ses systèmes automatisés posent des problèmes éthiques, comme nous l'avons déjà vu, l'IA formée sur des données de synthèse présente aujourd'hui de nouveaux scénarios inquiétants : comment garantir l'exactitude de la formation algorithmique ? Comment garantir la sécurité des systèmes mis en production s'ils n'ont jamais été réellement testés ?

Enfin, il convient de se demander si le consommateur ou l'utilisateur doit être informé de ces antécédents en matière d'IA : faut-il penser à une marque qui avertit le consommateur qu'il utilise un système qui a été formé ou qui est basé sur l'utilisation de données de synthèse ? S'agit-il d'une démocratisation de l'IA ou d'un terme commercial élégant qui cache le désir de produire des affaires avec des systèmes moins adaptés technologiquement ou plus fragiles ?

Le scénario se complique : pour gérer cette complexité, une philosophie algorithmique adéquate et des algorithmes appropriés sont de plus en plus nécessaires[7].

7. Voir le blog « paolobenanti.com », le 22.05.2018.

L'érosion de la réalité

Face à la puissance transformatrice et omniprésente de l'intelligence artificielle, de nombreuses voix, pas toutes technologiques, s'élèvent pour lancer des appels, pour noircir les scénarios ou pour demander des réglementations de ce nouveau et fascinant développement technologique. Toutes les voix ne s'accordent pas dans l'analyse du phénomène. Pour paraphraser Umberto Eco dans l'un de ses célèbres essais des années 1960, on pourrait dire qu'il existe de nombreuses tendances apocalyptiques et autant de tendances intégrées[8]. Parmi les apocalyptiques, on entend des alarmes sur la façon dont l'IA va mettre fin à notre société ou à l'espèce humaine elle-même ; beaucoup insistent sur l'avenir du travail et l'avènement des robots. Mais est-ce là le scénario à craindre ? Voyons si les transformations les plus radicales et les plus imminentes sont vraiment celles évoquées en premier.

Craindre la super-intelligence ?

Pour être clair dès le départ, je dois dire que je ne suis absolument pas convaincu que l'avènement de ce que beaucoup définissent comme une intelligence artificielle « super-intelligente » soit imminent ou constitue la menace la plus pressante des générations futures de machines fonctionnant selon des modèles d'apprentissage profond. En fait, pour être tout à fait franc, je ne suis pas du tout certain que toute la notion de super-intelligence soit réalisable et ne puisse être qu'une hypothèse philosophique, importante et sur laquelle il faut réfléchir. Nous ne savons pas si une telle IA pourra un jour être réalisée, développée ou évoluée dans le futur – ici sur terre ou ailleurs dans le cosmos. Cependant, bien que l'auteur ne se sente ni apocalyptique ni intégré, le développement d'une technologie aussi envahissante et transformatrice a des effets qui peuvent changer radicalement notre société et modifier tout aussi profondément les relations entre les humains et la compréhension que nous avons de nous-mêmes en tant qu'espèce. De mon point

8. Voir Umberto Eco, *Apocalittici e integrati. Comunicazioni di massa e teorie della cultura di massa*, 1964.

de vue, la transformation la plus proche et la plus radicale que l'IA peut produire est une distorsion radicale de ce que nous croyons être la vérité et des moyens que, en tant qu'hommes, nous partageons pour la rechercher.

À l'heure actuelle, dans cette confrontation avec les machines, que plusieurs voix définissent comme intelligentes, nous devons reconnaître que nous ne disposons pas d'une théorie quantitative convaincante de l'intelligence. Il n'y a pas de théorie qui nous dise ce que nous entendons par intelligence («regarde, il est intelligent, il peut ouvrir une boîte de haricots tout seul»), ni de théorie qui nous donne une échelle pour mesurer l'intelligence en la corrélant réellement avec la complexité, ni, enfin, s'il y a ou non un maximum théorique. Peut-être les réponses à ces questions ne peuvent-elles venir que d'une expérimentation adéquate ou du développement de cette théorie fondamentale de l'intelligence qui, comme nous l'avons dit précédemment, nous fait encore défaut.

Pour toutes ces raisons, je ne m'inquiète pas tant de l'avènement d'une IA super-intelligente sur terre que j'observe avec une attention scrupuleuse la propagation d'une IA relativement stupide dont le but – ou plutôt dont les capacités accessoires – est capable de manipuler notre relation avec l'information, avec ce que nous appelons les faits et avec la réalité telle que nous la percevons. Cette dimension de notre vie est peut-être la plus menacée par l'IA.

L'altération de la confiance

Pour comprendre cela nous devons introduire une petite digression sur le concept sociologique de la confiance et comment les sociétés et les croyances sont liées à ce concept.

La notion de confiance occupe une place qui est loin d'être secondaire dans la pensée politique et sociale occidentale. Les théories du contrat des XVII^e^ et XVIII^e^ siècles considéraient la confiance comme une condition essentielle de l'ordre politique et comme le fondement du contrat social. Même les pères fondateurs de la sociologie, plus intéressés à identifier l'élément moral qui imprègne l'ordre social, y font implicitement référence. Les contenus de la confiance systémique ou impersonnelle sont généralement qualifiés d'attentes de stabilité d'un ordre naturel et social donné, de reconfirmation donc

du fonctionnement de ses règles. Il s'agit donc d'attentes de régularité très diverses et généralisées.

Du point de vue cognitif, la confiance est placée dans une zone intermédiaire entre la connaissance complète et l'ignorance complète. L'attente de confiance intervient sur l'incertitude non pas en fournissant les informations manquantes, mais en les remplaçant par une forme de « certitude » interne qui a la valeur d'une réassurance positive en ce qui concerne les événements et les expériences éventuelles. L'incertitude est rendue plus tolérable par cet acte de substitution, qui réduit la complexité dans le sens de prédictions qui sont gratifiantes pour l'acteur. L'attente fiduciaire remplace donc l'incertitude par un niveau de « certitude » et d'assurance interne qui varie selon le degré de confiance accordé. Elle représente toutefois un investissement cognitif plus élevé que le simple espoir. En cas d'erreur, elle se heurte donc à des conséquences négatives plus graves en termes de motivation.

Les capacités de l'IA qui se répandent affectent précisément la formation, le développement et le maintien de la confiance individuelle et sociale. En utilisant des techniques telles que l'apprentissage contradictoire, nous avons déjà réalisé une IA qui peut imiter nos voix à la perfection. Des approches similaires pourraient sans doute être appliquées à notre style d'écriture, aux messages textuels et aux messages des médias sociaux. En usurpant notre apparence par l'analyse de photos, nous pouvons produire de fausses photos ou générer des vidéos où nous faisons apparemment des choses que nous n'avons jamais faites auparavant.

Ces systèmes iront probablement plus loin (si ce n'est déjà fait, il est difficile de suivre l'évolution). Pourquoi ne pas générer des nouvelles entières ou des colonnes de ragots avec une IA ? Les tabloïdes hollywoodiens n'exigent presque jamais de faits, et les nouvelles grand public semblent parfois suivre le mouvement.

Dans l'IA, il existe un potentiel extraordinaire pour générer des flux de communication qui peuvent tromper quelqu'un. Ce peut être par exemple voler nos données personnelles en nous trompant, ou créer une version alternative de nous qui se livre à tout type d'acte antisocial, voire criminel. Ou ce peut être encore en nous manipulant pour que nous voulions certains biens, ou que nous votions d'une certaine manière, ou que nous croyions en certaines choses. Si nous devions développer la première IA évangélique, cela pourrait

facilement surpasser même les meilleurs prédicateurs humains. Et contrairement à la super-intelligence hypothétique (dont les motivations sont difficiles à imaginer), l'utilisation de l'IA pour exploiter les gens ou pour forcer la société suit un modèle politique très ancien.

Les capacités humaines de collaboration

Nous, les humains, avons probablement érodé la réalité depuis le moment où nos ancêtres *Homo sapiens* ont commencé à communiquer et à raconter des histoires. Une bonne narration autour du feu peut aider à maintenir une histoire orale, ou à articuler des règles morales et sociales, contribuant à apporter de la cohésion à nos familles et à nos groupes. Mais elle peut, peut-être inévitablement, induire en erreur, déformer et manipuler.

Suivons, en le paraphrasant, le raisonnement sur le thème que Yuval Noah Harari développe[9]. Il y a 70 000 ans, nos ancêtres étaient des animaux insignifiants. La chose la plus importante à savoir sur l'homme préhistorique est qu'il n'était pas important. Son impact sur le monde n'était pas beaucoup plus important que celui des méduses. Aujourd'hui, au contraire, nous dirigeons cette planète. La question est de savoir comment nous avons fait. Comment sommes-nous passés de singes insignifiants, s'occupant de leurs propres affaires dans un coin de l'Afrique, à des dirigeants de la planète Terre ?

Les humains contrôlent la planète parce qu'ils sont les seuls animaux capables de collaborer avec souplesse et en grandes masses. Certes, il existe d'autres animaux, comme les insectes sociaux – abeilles, fourmis – qui peuvent collaborer en grand nombre, mais ils ne collaborent pas de manière flexible. Ils ne collaborent que sous des formes strictement prédéfinies. D'autres animaux, tels que les mammifères sociaux, les loups, les éléphants, les dauphins, les chimpanzés, sont capables de collaborer de manière beaucoup plus souple, mais ils ne le font qu'en petit nombre, car la collaboration entre chimpanzés est basée sur une connaissance mutuelle intime.

Le seul animal capable de combiner ces deux compétences, c'est-à-dire de collaborer avec souplesse et de le faire même en grand nombre, est l'*Homo sapiens*. Mais comment faire exactement ? Qu'est-ce qui

9. Voir Yuval Noah Harai, *Sapiens. Une brève histoire de l'humanité*, Albin Michel, 2015.

nous permet, nous qui sommes uniques parmi les animaux, de collaborer de cette manière ? La réponse est notre imagination. Nous sommes capables de collaborer de manière flexible et avec un nombre infini d'étrangers parce que nous sommes les seuls, parmi tous les animaux de la planète, à pouvoir créer et croire à des fictions, à des histoires imaginaires. Si tout le monde croit à la même fiction, alors tout le monde obéit et suit les mêmes règles, les mêmes normes, les mêmes valeurs.

D'un point de vue biologique, cependant, altérer la réalité est un comportement qui ne se limite pas à notre espèce « intelligente ». La tromperie est très répandue dans le monde naturel. Les animaux se camouflent, ou font semblant d'être des choses qu'ils ne sont pas – de l'imitation de l'apparence d'espèces venimeuses au gonflage des plumes, des écailles ou de la peau ; les mâles de nombreuses espèces se parent ou construisent des structures séduisantes et recourent à de profonds subterfuges pour tenter de propager leurs gènes. La tromperie semble faire partie de la sélection darwinienne au même titre que l'honnêteté. La capacité à induire en erreur est une mesure de l'aptitude à l'évolution.

Nous ne devrions pas avoir de grands espoirs d'obtenir des machines qualitativement différentes des modèles sur lesquels elles sont formées : si nous créons des artefacts qui ont pour seule caractéristique, par conception, l'optimisation du résultat ou la victoire dans un jeu qui simule la situation de prise de décision, peut-être devrions-nous nous interroger d'urgence sur les conséquences sociales que la diffusion de tels systèmes peut avoir dans le tissu social hyper-connecté que nous habitons.

Bien sûr, comme l'a dit le physicien Niels Bohr, il est terriblement difficile de faire de bonnes prédictions, surtout lorsqu'il s'agit de l'avenir. Mais une chose est sûre, nous en apprendrons beaucoup sur la trajectoire que peut suivre l'intelligence – en supposant, bien sûr, que nous puissions voir, connaître et dire la vérité[10].

Paolo Benanti
Université Grégorienne, Rome
Académie Pontificale pour la Vie
(traduit de l'italien par Alain Thomasset, avec l'aide de l'IA DeepL)

10. Voir le blog « paolobenanti.com », le 03.07.2018.