Enjeux numériques

Intelligences artificielles et humaines, quelles interactions ?

UNE SÉRIE DES
ANNALES DES MINES
FONDÉES EN 1794

N° 12 - DÉCEMBRE 2020

Publiées avec le soutien de l'Institut MinesTélécom

UNE SÉRIE DES
ANNALES DES MINES
FONDÉES EN 1794

ENJEUX NUMÉRIQUES

Série trimestrielle • N°12 - Décembre 2020

Rédaction
Conseil général de l'Économie,
ministère de l'Économie, des Finances et de la Relance
120, rue de Bercy - Télédoc 797
75572 PARIS Cedex 12
Tél. : 01 53 18 52 68
http://www.annales.org

François Valérian
Rédacteur en chef

Gérard Comby
Secrétaire général

Alexia Kappelmann
Secrétaire générale adjointe

Magali Gimon
Assistante de rédaction

Myriam Michaux
Webmestre et maquettiste

Membres du Comité de rédaction

Jean-Pierre Dardayrol
Président du Comité de rédaction

Edmond Baranes
Godefroy Beauvallet
Côme Berbain
Pierre Bonis
Serge Catoire
Michel Cosnard
Arnaud de La Fortelle
Caroline Le Boucher
Alban de Nervaux
Bertrand Pailhès
Grégoire Postel-Vinay
Jacques Serris
Hélène Serveille
Laurent Toutain
Françoise Trassoudaine
François Valérian

Photo de couverture :
Robert Delaunay (1885-1941), *Rythme n°2*, huile sur toile, 1938.
Paris, musée d'Art moderne.
Photo ©Musée d'Art moderne/Roger-Viollet.

Iconographie
Christine de Coninck

Abonnements et ventes
COM & COM
Bâtiment Copernic - 20, avenue Édouard-Herriot
92350 LE PLESSIS-ROBINSON
Alain Bruel
Tél. : 01 40 94 22 22 - Fax : 01 40 94 22 32
a.bruel@cometcom.fr

Mise en page : Nadine Namer
Impression : EspaceGrafic
N° ISSN : 2607-9984

Éditeur délégué
FFE – 15, rue des Sablons - 75116 PARIS
www.ffe.fr

Régie publicitaire : Belvédère Com
Fabrication : Aïda Pereira
aida.pereira@belvederecom.fr
Tél. : 01 53 36 20 46
Directeur de la publicité : Bruno Slama
Tél. : 01 40 09 66 17
bruno.slama@belvederecom.fr

Le sigle « D. R. » en regard de certaines illustrations correspond à des documents ou photographies pour lesquels nos recherches d'ayants droit ou d'héritiers se sont avérées infructueuses.

Intelligences artificielles et humaines, quelles interactions ?

Les mutations engendrées par les IA : puisque les IA vont nous permettre de faire plus de choses, comment allons-nous nous y adapter (et réciproquement) ?

Ce numéro a été coordonné par Arnaud de LA FORTELLE.

Introduction : Intelligences artificielles et humaines, quelles interactions ?

Par **Arnaud de LA FORTELLE**
PSL – MINES ParisTech

L'intelligence artificielle fascine : l'esprit humain démultiplié par la machine ! Les conséquences sont fantastiques, en termes opérationnels, mais aussi sociaux, à tel point que le développement des intelligences artificielles (IA) a suscité des mises en garde contre une possible fin de l'humanité par des personnes comme Stephen Hawking ou Bill Gates. Bien d'autres personnes ont appelé à la vigilance, il n'est que de lire les signataires de la lettre ouverte de l'institut Future of Life [(1)] : il ne s'agit pourtant pas de technophobes ni d'ignares. On peut donc se poser la question : où allons-nous ?

En premier lieu, qu'est-ce ? Une présentation classique de l'intelligence artificielle est de la comparer aux activités humaines, la plupart du temps afin de les mettre en confrontation (« test de Turing »), voire en concurrence, parfois jusqu'au point – pour les tenants du transhumanisme – où l'humain serait totalement supplanté. C'est une présentation « efficace », attractive, mais finalement très prospective et éloignée d'aujourd'hui. Alors oui, les conséquences du déploiement des algorithmes d'IA sont impressionnantes, mais la réalité ne ressemble pas aux œuvres de fiction. C'est tout l'objet de ce numéro, se confronter à ce que fait l'intelligence artificielle, et comment elle le fait. Mais aussi se poser la question de ce qu'elle ne fait pas. Aussi avons-nous demandé à quelques experts de nous apporter leurs lumières sur cette question : quelles interactions entre intelligences artificielles et humaines ? Collaboration, confrontation ou substitution ? Et adaptation ?

Une première partie de ce numéro présente des cas concrets d'application de l'intelligence artificielle, vue comme des algorithmes d'inférences statistiques (apprentissage) qui permettent d'effectuer certaines tâches : reconnaissance d'objets, traitement d'information, prédiction, etc. Il existe de nombreux problèmes concrets pour lesquels ces techniques apportent dès aujourd'hui de réels bénéfices pour tous : « La "cobotique" et l'interaction homme-robot » et « L'intelligence artificielle en milieu industriel, levier de transformation et facteur d'innovation du groupe RATP » en sont de bonnes illustrations. Là se pose déjà la question de l'interaction entre les experts humains et les « performances surhumaines » des IA sur certaines tâches, comme le décrit l'article de Valeo sur la conduite autonome. Même pour des tâches comme la compilation de quantités de données inaccessibles à l'homme, appliquées à la stratégie d'entreprise (article de M. Isaac), on voit bien que l'IA apporte des analyses très appréciables, mais que sa production n'est pas parfaite, et que le déploiement de telles capacités nécessite des ajustements. On le constate vite : l'IA est en phase de déploiement massif. Cependant, même si sa puissance est telle qu'aucune activité ne semble pouvoir y échapper, le retour d'expérience que nous en avons montre qu'elle ne peut pas non plus tout faire.

Les processus actuellement déployés commercialement ou en passe de l'être sont toujours en interaction avec des processus humains. La première partie de ce numéro introduit une interaction avec des experts, *data scientists* ou ingénieurs. Qu'en est-il de tous les autres ? Devrons-nous nous plier aux décisions des IA, plus rapides, capables de digérer des montagnes de données, et que seuls quelques rares experts seraient capables de comprendre ? C'est la question difficile

(1) https://futureoflife.org/ai-open-letter

de l'interprétabilité des IA, que Mme Guégan présente dans son article. On y voit apparaitre des notions telles que le biais des données, l'équité et le risque : produire des IA qui prennent en compte ces éléments nous force à les préciser. Quelle équité voulons-nous ? Pour les IA comme pour les humains, c'est une question difficile. Car les données, ou les nombres, ne sont pas une vérité absolue. L'article « Intelligence artificielle et contrôle de gestion : un rapport aux chiffres revisité et des enjeux organisationnels » souligne la question du sens que nous donnons à des indicateurs, et de la construction de ce sens. Il y a bien un risque à l'utilisation des IA, qui est un risque à maîtriser, comme pour la sécurité (article de M. Barnu). Cependant, le risque peut devenir réalité, et le domaine du droit a également besoin de s'ajuster : « Quelle régulation juridique pour l'intelligence artificielle ? ». Dès aujourd'hui, les IA nous renvoient à des questions profondes sur nos organisations et sur nos objectifs : c'est une forme d'interaction, qui n'est pas un réel dialogue, mais qui va certainement continuer à bousculer nos manières de faire.

Nous avons désormais des IA qui prennent place dans nos organisations. Mais nous le voyons aussi, elles évoluent très vite, ou plutôt la technique évolue très vite : en effet, nous n'avons encore jamais vu d'IA évoluer en dehors de son champ : le champion du jeu de Go serait probablement bien incapable de conduire une voiture. Pouvons-nous brosser un tableau de ce que pourraient être les évolutions des IA d'ici quelques années ? La prospective tourne assez facilement à la divination, et nous nous sommes limités à un horizon temporel plutôt court, de quelques années, où nous espérons avoir une visibilité raisonnable. Il est donc intéressant de s'intéresser à nos représentations des IA (« Une IA ou des IA ? ») afin de mieux comprendre ce que pourraient être nos relations avec elles. Ensuite, et c'est une question brûlante, les IA prendront-elles nos emplois ? L'article « Intelligence artificielle et travail : le défi organisationnel » décrit bien la mutation en cours, avec des remplacements, certes, mais aussi beaucoup de complémentarité et surtout les importants ajustements nécessaires, pour les travailleurs comme pour les organisations, afin d'en tirer parti. Dans « Le futur du travail en présence de formes artificielles d'intelligence », qui offre une vision tout à fait complémentaire (les auteurs des deux articles ayant livré des analyses indépendantes), un lien très fort entre l'homme et les IA est dessiné, avec des opportunités (qui peuvent être des risques) pour la société : saurons-nous tirer le meilleur parti de ces nouvelles techniques ? L'application des techniques d'IA au droit (« Algorithmes et droit pénal : quel avenir ? ») démontre l'importance et les ramifications des adaptations à mener. Par ailleurs, la gestion des émotions – bien humaines – par les IA (les « machines empathiques » de Mme Devillers) va transformer notre relation, et le rapprochement de cet article avec le précédent, avec des IA « gérant » le droit et les émotions, ne peut que nous interpeller. Sans doute le processus à l'œuvre avec les IA fait-il partie des transformations que l'humanité a traversées pour devenir ce qu'elle est aujourd'hui. Car si les IA peuvent évoluer, nous le savons, nous en sommes capables aussi.

En conclusion, les intelligences artificielles et humaines se répondent bien : les IA ont besoin de nous, à la fois pour les développer, mais aussi pour qu'elles soient efficaces dans leurs tâches et dans nos organisations. Elles sont déjà en train de transformer notre monde, dans lequel elles sont en train de trouver leur place, peut-être comme tous les outils que nous avons créés, peut-être un peu plus. Nous espérons que ce numéro permettra à chacun de s'en faire une opinion.

Bonne lecture !

L'intelligence artificielle en milieu industriel, levier de transformation et facteur d'innovation du groupe RATP

Par **Côme BERBAIN**
Directeur de l'Innovation du groupe RATP
Et **Yohan AMSTERDAMER**
Responsable du « programme intelligence artificielle » du groupe RATP

Le groupe RATP conçoit, exploite et maintient des modes de transport variés (bus, tram, métro, RER, bateau, câble) répondant aux besoins de mobilité des voyageurs, en Île-de-France, mais également dans d'autres villes en France et à l'international.

Face à une adaptation longue des infrastructures industrielles dans le temps et à une demande toujours plus forte de transport, le groupe RATP se doit d'être capable d'actionner tous les leviers technologiques comme le numérique et l'intelligence artificielle afin de permettre une amélioration aussi bien de l'excellence opérationnelle que de l'expérience client.

Un programme d'innovation au service des clients et des acteurs du transport

Dans un contexte où la France et l'Europe souhaitent se positionner comme des acteurs majeurs de l'intelligence artificielle (IA), notamment dans le secteur des transports et de la mobilité, le groupe RATP a engagé un programme stratégique pour faire de l'IA un des piliers de sa transformation numérique.

L'histoire de « l'intelligence machine » à la RATP est ancienne : on peut la retracer dans un *continuum* qui va du contrôle automatique des aiguillages aux premiers systèmes d'aide à la conduite, des systèmes automatiques de vente aux capteurs équipant les infrastructures et les matériels roulants, jusqu'aux systèmes permettant la planification et l'optimisation de la production de l'offre de transport.

La création de lignes de métro automatiques puis l'automatisation de lignes de métro, tout en maintenant le service, sont devenues des marqueurs des capacités du Groupe et ont été l'occasion de repenser l'ensemble de la chaîne opérationnelle : de la conception à l'exploitant, jusqu'au mainteneur et l'expérience des voyageurs. Pour cela, la RATP s'est appuyée sur la recherche opérationnelle et un premier type d'IA : l'IA déterministe ou « symbolique ».

Aujourd'hui, la valeur potentielle des données, l'amélioration de la performance des algorithmes et le développement de la puissance informatique permettent d'ouvrir et d'explorer de nouveaux champs d'application : c'est l'ère de l'IA statistique ou « connexionniste ». Dans le milieu industriel, cette IA ouvre la voie vers la modélisation de nouvelles classes de phénomènes, une théorisation plus rapide et locale des plus complexes d'entre eux, voire l'augmentation de la performance des modèles déterministes existants.

Ainsi, un « programme intelligence artificielle » a été initié et structuré en 2018 afin d'adresser ces enjeux autour de plusieurs axes définis à partir des besoins métiers des collaborateurs et l'analyse des besoins de nos clients, conformément aux valeurs du Groupe de responsabilité et d'inclusion.

Quatre axes ont été priorisés :

- Augmenter la performance de la production industrielle de l'offre de transport :
  - par une appréhension plus fine et en temps réel des phénomènes de perturbations du trafic et des imprévus d'exploitation ;
  - par une modélisation en temps réel, à la maille du réseau multimodal, du besoin de mobilité et des flux de voyageurs pour optimiser l'allocation des moyens de production.
- Assurer une maintenance plus proactive et efficiente de l'appareil industriel afin d'améliorer la disponibilité et le fonctionnement des équipements :
  - par la consolidation de connaissances contextualisées et locales du comportement des équipements du réseau (par exemple : matériels roulants, infrastructures, espaces d'accueil voyageurs, dispositifs d'information voyageurs, etc.) ;
  - par l'identification de corrélations entre causes, conséquences, symptômes de défaillances et opérations de maintenance.
- Agir sur l'expérience client en proposant un voyage engageant et personnalisé, adapté aux différents usages :
  - par la compréhension des caractéristiques, des attentes et des besoins clients à chaque étape de leur parcours, notamment au travers de l'information voyageur ;
  - par une capacité à outiller les acteurs du service client du Groupe, pour davantage de satisfaction personnelle et d'efficience.
- Améliorer le sentiment de sécurité pour nos voyageurs, nos collaborateurs et nos sites industriels afin de détecter, qualifier et prévenir les risques :
  - par la valorisation des données vidéo en temps réel issues du réseau de vidéoprotection.

De la connaissance des experts vers des outils et services au bénéfice des voyageurs et des collaborateurs

Le savoir-faire et la connaissance des experts du transport ont été informatisés progressivement. Ceci a permis d'automatiser des processus « métier » comme superviser et réguler une ligne de métro, des tâches complexes comme la construction des horaires et des plannings des conducteurs, et également des tâches simples, mais répétitives, comme la vente de titres de transport. Cette informatisation a permis d'accompagner la hausse de la demande de mobilité et également la satisfaction des clients et des collaborateurs.

Cette connaissance d'expert est développée, acquise et maintenue par un long processus d'essais/erreurs, de transmissions, d'observations, d'expériences,... qui ont permis une théorisation experte solide et maitrisée. Tout l'enjeu a été de « traduire » cette connaissance vers la machine pour qu'elle puisse l'exécuter. Le Groupe a mis en œuvre ce principe de théorisation à maintes reprises sur des phénomènes déterminés.

Mais comment modéliser et décrire les phénomènes aléatoires et statistiques comme les perturbations de trafic, la survenue d'un incident, les usages et les besoins voyageurs en évolution continue ?

C'est sur ces sujets que l'IA vient proposer des réponses adaptées en s'appuyant sur des données en grand nombre et en permettant la modélisation de nouveaux phénomènes. Tout l'enjeu réside dans la manière de produire ces données, de les acquérir, de les qualifier, de les comprendre et de leur donner du sens pour en permettre une interprétation et une analyse opérationnelle. Le rôle de l'expert et la connaissance métier et terrain sont encore une fois au centre. Cette expertise humaine intervient sur l'ensemble du cycle de vie des données : de la conception et l'ingénierie

à l'exploitation et la maintenance des infrastructures de transport et des systèmes informatiques industriels. Ce sont ces mêmes experts qui sont à la genèse, la production et l'interprétation de ces données, qui valident le comportement et les sorties de ces modèles : c'est l'expert qui apprend à l'IA en fin de compte, avec l'appui d'un nouvel expert des mathématiques et de l'informatique, le *datascientist*.

Ce processus permet, au passage, de formaliser et de capitaliser une connaissance métier parfois cantonnée à l'expertise individuelle issue de l'expérience personnelle, ce qui facilite la transmission de connaissances et la performance collective. L'un des premiers projets du programme IA au service de la régulation du trafic sur le RER B en est un exemple emblématique. Face à une perturbation du trafic donnée, les actions de régulation entreprises par les aiguilleurs et les chefs de régulation reposent sur un cadre de sécurité ferroviaire, mais aussi sur leurs expériences et vécus personnels. L'analyse que ces experts métiers font de la situation et la décision qu'ils prennent comme étant la « meilleure possible » relèvent de connaissances peu formalisées et transmises par la pratique. Confronter ces experts métiers à l'IA a permis de formaliser, théoriser, modéliser et généraliser ces « bonnes pratiques ».

Les systèmes déterministes permettent donc de représenter, décider, planifier, optimiser alors que l'IA, elle, permet d'analyser et de prévoir à partir des données. C'est cette combinaison des systèmes déterministes existants et de l'IA probabiliste qui va permettre de construire les outils d'aide à la décision, à l'exploitation et à la maintenance de demain, ainsi que d'améliorer la qualité de service et l'expérience des clients de la mobilité : le secteur industriel a besoin de l'ensemble de la chaîne de fonctions.

Le secteur industriel implique donc la coexistence de plus en plus importante de ces deux mondes de l'IA et ils ne pourront être dissociés. Le premier monde fait intervenir un expert ingénieur alors que le second introduit comme acteur incontournable l'expert métier. C'est ce binôme d'experts qui devient l'entraineur de l'IA du second monde.

De l'intelligence des données vers la connaissance des experts

L'informatisation et l'automatisation massives du secteur des transports et de la mobilité a pendant ce même temps établi des systèmes de plus en plus complexes et générant de plus en plus de données, qui sont autant de potentiels pour l'IA d'approcher de nouvelles classes de phénomènes que l'expert ne peut rendre intelligibles avec ses connaissances existantes.

En effet, certains espaces de connaissance experte sont restés incomplets, inexplorés ou sous exploités par manque de robustesse ou de théorie suffisante, ou encore non accessibles par manque d'existence, d'accès ou de qualité des données. La capacité à prédire un retard sur une ligne ferroviaire en fonction du trafic en est un premier exemple. L'identification fine des précurseurs et des modes de défaillance des systèmes de nouvelle génération en est un autre. De plus, la corrélation entre ces défaillances et le contexte d'exploitation et d'opérations de maintenance va aussi accélérer la constitution de cette nouvelle connaissance. Pour prendre un dernier exemple, on peut citer la faculté à modéliser et anticiper au plus proche du terrain en temps réel les flux de voyageurs en fonction de l'offre ; la période de crise sanitaire l'a réaffirmé.

L'IA pose néanmoins des questions de formalisation et d'intelligibilité de cette nouvelle connaissance. Mais l'important est-il toujours de théoriser ou de savoir interpréter et déclencher les meilleures actions ? Si l'IA peut établir finement les classes d'incidents des équipements industriels, l'expert métier, lui, n'a pas besoin de comprendre le modèle mais d'appréhender les variables explicatives prépondérantes qui vont lui permettre de déclencher les bonnes opérations de maintenance au bon moment et d'agir à moyen terme sur son ingénierie de maintenance.

Le projet au service de la régulation du trafic sur le RER B déjà évoqué en est un exemple. La constitution du modèle de perturbation s'appuie sur des variables explicites comme l'écart à l'heure de pointe, la typologie de l'offre de transport en cours…

Cette capacité d'analyse locale et profonde permet donc de créer de la connaissance dans certains domaines encore peu explorés, répondant de complexités nouvelles, ou d'aller encore plus loin. L'IA statistique permet ainsi, dans certains cas, d'accélérer la modélisation de ces phénomènes. Mais cette connaissance est générée sous une forme nouvelle qui implique un changement de culture important. L'ingénierie des données constitue le champ d'expertise révélateur de ce changement de conception et de structuration de la connaissance. La difficulté relève de la dualité que cette expertise nécessite : le binôme ingénieur et métier. Sa mise en pratique en entreprise va contraindre à repenser les parcours de carrière, d'une part, et accélérer la mise en pratique de la conduite de projets agiles, d'autre part, afin de rompre les frontières entre prescripteurs, réalisateurs et utilisateurs.

Ainsi, la complexité croissante des systèmes et le besoin continu de prendre en compte le contexte associé aux usages rendent incontournables ces nouveaux gisements de connaissances. À celles-ci de nous informer, de nous apprendre, pour atteindre de nouvelles performances et de nouveaux services ! C'est le juste retour que les données et l'IA vont fournir à l'expert !

L'intelligence artificielle, un processus apprenant favorisant l'innovation

L'intelligence artificielle introduit une pensée nouvelle sur les processus d'apprentissage de l'entreprise. Elle propose un cercle vertueux entre l'apprentissage que peut réaliser l'expert vers les machines et le juste retour que les données peuvent lui apporter. Dans un sens, cela permet d'automatiser certaines tâches et opérations pour gagner en efficacité et en service dans un secteur de la mobilité où la demande et l'exigence sont grandissantes. Et dans l'autre, cela permet d'accéder à de nouvelles fonctions et de nouvelles connaissances qui vont servir l'expert, les opérateurs et les clients.

Dans les faits, c'est une dynamique qui s'opère dans un processus apprenant continu, qui bénéficie à l'excellence opérationnelle et à l'expérience client du secteur du transport. Cela fait repenser le processus de conception des systèmes d'information ainsi que la chaîne opérationnelle, pour passer d'un modèle statique et déterministe à un modèle apprenant et probabiliste.

Le monde industriel comporte tout de même un facteur limitant : l'accès, la qualité, la qualification, voire même l'existence de la donnée et de son historique. Cela nécessite un changement dans l'ingénierie des données et des systèmes d'information. Cette transformation va être progressive, néanmoins, l'existence de la donnée n'est pas toujours de l'ordre du possible ni la connaissance experte toujours en expansion. Par exemple, disposer d'un nombre suffisant d'images 2D de chaque catégorie de défauts de *boggie* pour spécialiser un dispositif de reconnaissance automatique n'est pas réaliste ! Il faut donc toujours établir le bon curseur entre l'IA déterministe et l'IA statistique afin d'apporter la meilleure solution à la problématique posée. Et même lorsqu'elles existent, comme les données vidéo et la vision artificielle, il n'est pas toujours simple, voire souhaitable de s'en servir.

Un processus apprenant favorise et stimule l'innovation ! Plus les connaissances expertes et celles de la machine s'étendent, plus les idées, les besoins et les concepts innovants émergent.

La « cobotique » et l'interaction homme-robot

Par **Vincent WEISTROFFER**
Ingénieur-chercheur au CEA-List dans le Laboratoire de Simulation Interactive

La « cobotique » dans l'industrie

Dans l'industrie, l'apparition des robots s'est faite progressivement pour prendre en charge des postes pénibles ou dangereux. C'est le cas dans l'industrie automobile, dans laquelle une partie importante des chaînes de montage est automatisée : les postes de ferraillage sont réalisés par des robots imposants pour souder les parties métalliques entre elles, tandis que les postes de peinture sont réalisés par des robots spécialement conçus pour résister aux vapeurs toxiques. Ces robots sont complètement autonomes et interdisent l'intervention de l'homme dans leur espace de travail pour des raisons de sécurité.

Néanmoins, il n'est pas envisageable de robotiser la totalité des postes industriels. Certains postes nécessitent l'agilité, la flexibilité, l'adaptabilité et l'intelligence d'opérateurs humains. C'est par exemple le cas des postes de montage automobile, dans lesquels les opérateurs assemblent de petites pièces dans des endroits parfois difficiles d'accès, tout en gardant un temps de cycle performant et en s'adaptant aux aléas. Pour certains postes, les opérateurs sont assistés par des manipulateurs qui permettent de porter des charges lourdes (comme un tableau de bord), mais qui restent guidés par l'humain.

Pour des raisons d'ergonomie, de performance et de coût, il peut être intéressant de coupler les compétences des opérateurs à celles des robots sur un même poste de travail. Pour permettre aux opérateurs et aux robots de travailler ensemble dans les mêmes espaces, il est nécessaire de mettre en place de nouveaux types de robots, souvent appelés « cobots » (pour robots collaboratifs). Ces

Figure 1 : Montage de porte en « co-présence » avec un robot (Weistroffer, 2014).

« cobots » sont intrinsèquement conçus pour garantir la sécurité des opérateurs à proximité. Ils permettent alors le partage d'espace avec des opérateurs (on parle de « co-présence »), le partage d'activités pour des opérations successives (on parle de « co-action ») ou encore le partage de tâches communes (on parle de « collaboration » ou de « co-manipulation »).

Aujourd'hui, l'utilisation des « cobots » dans l'industrie est soumise à des certifications et des normes de sécurité, comme la norme ISO 15066. Ces normes spécifient des critères de sécurité liés à la vitesse du robot, à l'énergie et aux efforts déployés, ainsi qu'à la distance aux opérateurs. L'utilisation des « cobots » est aujourd'hui plutôt déployée dans des contextes de « co-présence », mais leur utilité dans des contextes plus collaboratifs reste étudiée par les industriels. Notamment, les « cobots » représentent une solution intéressante pour les PME, dont les processus de fabrication doivent être de plus en plus flexibles pour s'adapter aux besoins des clients.

Contrairement à la robotique industrielle classique, la « cobotique » fait intervenir l'humain dans la boucle, en permettant de partager les espaces de travail et d'interagir ensemble. De nouvelles problématiques de recherche apparaissent, autour de la sécurité, de l'ergonomie du poste et de la performance. C'est dans ce contexte que les techniques d'apprentissage artificiel peuvent amener des pistes de réflexion intéressantes.

L'apprentissage artificiel pour la « cobotique »

Avec l'émergence de la « cobotique », les industriels doivent faire face à de nouvelles problématiques de développement. L'une de ces problématiques concerne la transition des postes existants vers des postes collaboratifs et l'aide à la conception de ces nouveaux postes. Une autre problématique concerne l'interaction entre l'homme et le robot et la meilleure façon de collaborer ensemble. Dans la suite, nous abordons ces deux problématiques indépendamment et essayons d'illustrer, pour chacune d'elles, comment peut intervenir l'apprentissage artificiel pour faciliter leur mise en œuvre.

L'aide à la conception du poste collaboratif

La transition d'une robotique industrielle classique à une robotique plus collaborative nécessite, de la part des industriels, de réfléchir à une nouvelle configuration des postes de travail. Alors que, sur les postes traditionnels, les robots sont enfermés derrière des barrières physiques, les « cobots » autorisent désormais un partage d'espace avec les opérateurs. Même si les « cobots » sont dits « intrinsèquement sûrs » (c'est-à-dire qu'ils peuvent détecter les collisions avec les opérateurs et s'arrêter en conséquence, grâce à des capteurs embarqués), des capteurs externes sont parfois nécessaires pour respecter les contraintes imposées par les certifications de sécurité.

Les nouveaux postes collaboratifs sont ainsi plus complexes à concevoir, car un grand nombre de paramètres entrent en jeu. Les outils traditionnels de modélisation rencontrent aujourd'hui des limitations dans la prise en compte des paramètres dynamiques du robot et dans le calcul précis de critères liés à la sécurité, comme la masse apparente ou les efforts déployés par l'effecteur du robot.

Le projet SEEROB [1] (Simulation Ergonomique des Environnements de travail avec des ROBots collaboratifs), développé au CEA-List dans la communauté FactoryLab [2] avec PSA et Safran, propose un outil d'aide à la conception des postes collaboratifs. Cet outil permet de visualiser le futur poste collaboratif, de simuler le comportement dynamique des robots, de choisir différents capteurs de sécurité et de calculer en temps réel un ensemble de critères liés à la sécurité et à

(1) https://www.youtube.com/watch?v=9wMAvunW5jM
(2) http://factorylab.fr/

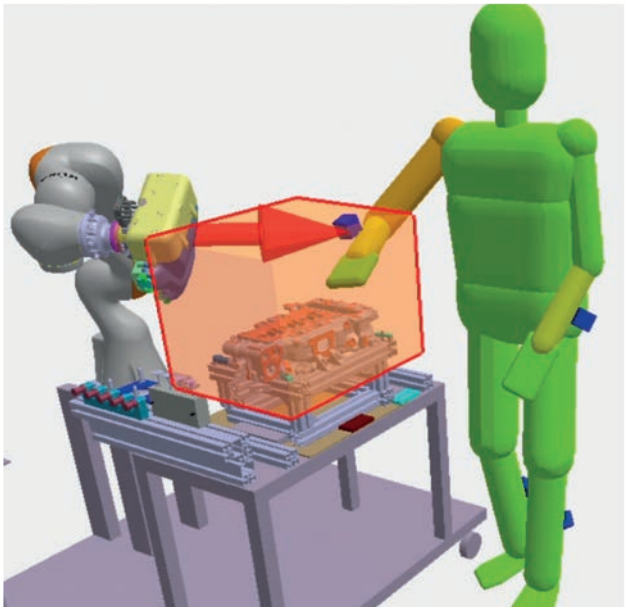


Figure 2 : Simulation dans le logiciel SEEROB pour l'analyse de l'ergonomie et de la sécurité des postes de travail avec des robots collaboratifs.

l'ergonomie du poste. Cet outil peut également être utilisé sur un poste déjà existant dans l'optique de générer un rapport de certification. Cet outil fait aussi intervenir l'opérateur dans la boucle, en utilisant des techniques de réalité virtuelle pour immerger l'opérateur dans la simulation à échelle 1, à l'aide de casques de réalité virtuelle ou de réalité mixte.

Les ingénieurs en robotique disposent ainsi de certains outils d'aide à la conception du poste collaboratif. Cette conception peut néanmoins se révéler laborieuse dans le cas où beaucoup de paramètres sont à prendre en compte : une bonne expertise peut être nécessaire pour optimiser le placement d'un robot ou la configuration de capteurs de sécurité. C'est dans ce contexte que l'apprentissage artificiel peut apporter son soutien, notamment avec des techniques d'apprentissage par renforcement : l'agent autonome choisit un jeu de paramètres et la simulation lui retourne une récompense (plusieurs scores de sécurité par exemple) ; après plusieurs expériences itérées, l'agent optimise au fur et à mesure son jeu de paramètres. Ces techniques sont encore plus efficaces lorsque plusieurs simulations peuvent être lancées en parallèle sur des *clusters* d'ordinateurs. À l'issue des simulations, les meilleurs résultats sont sélectionnés et placés dans un panel de solutions candidates pour un futur poste collaboratif.

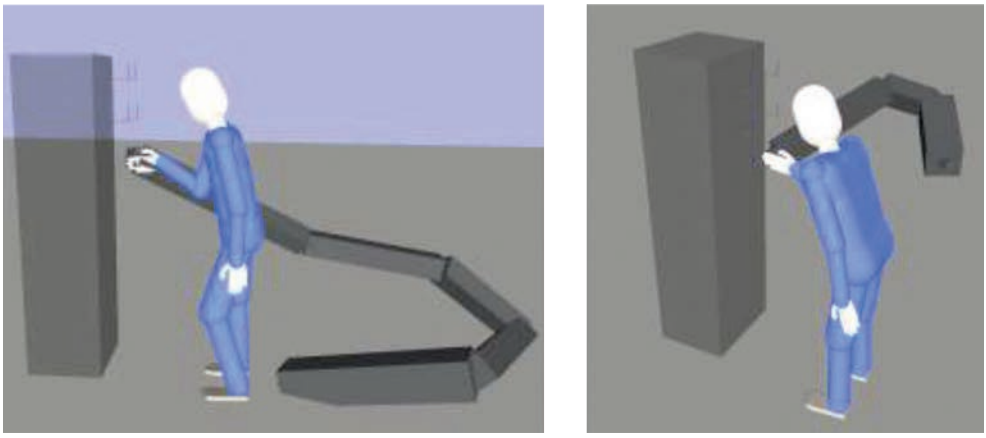


Figure 3 : Utilisation d'algorithmes génétiques dans la conception d'un « cobot » sur une tâche de perçage. À gauche : 10^{e} génération. À droite : 220^{e} génération. Au fur et à mesure des générations, la taille du robot décroît tandis que les sollicitations biomécaniques du mannequin virtuel sont améliorées (Maurice, 2015).

Ces techniques de simulation hors ligne sont utilisées en robotique dans d'autres contextes, par exemple dans la génération de trajectoires ou la génération de prises. Ces techniques ont également été utilisées par Maurice (2015), afin de déterminer la configuration optimale d'un « cobot » (position, structure cinématique) pour soulager les troubles musculo-squelettiques de l'opérateur sur une tâche de perçage. Des algorithmes génétiques ont ainsi été utilisés pour simuler un grand nombre de configurations et converger vers une solution optimale.

L'interaction homme-robot

Une fois qu'un poste collaboratif est conçu au niveau macroscopique, pour garantir sa sécurité et son ergonomie il convient de se pencher plus précisément sur l'interaction entre l'homme et le robot, et la manière dont ils peuvent collaborer. Avec la robotique industrielle classique, les robots répètent les mêmes cycles d'opérations de manière déterministe et n'engagent pas d'interaction avec les opérateurs. Avec le partage d'espace rendu possible par la « cobotique », les robots doivent s'adapter au comportement des opérateurs à proximité, que ce soit pour éviter une collision ou pour échanger des objets. La programmation des robots devient ainsi plus complexe et il devient nécessaire de les doter d'une intelligence supplémentaire pour reconnaître l'activité de l'opérateur et s'adapter en conséquence.

Pour certaines configurations de poste, tout contact entre l'homme et le robot est interdit par mesure de sécurité ; pour d'autres configurations, le contact est nécessaire, car l'opérateur et le robot doivent travailler ensemble sur les mêmes pièces, ou alors l'opérateur doit guider le robot à des endroits précis. L'un des enjeux de la robotique collaborative consiste non seulement à détecter les collisions entre l'homme et le robot, mais également à classifier ces collisions pour distinguer les contacts intentionnels des contacts accidentels, potentiellement dangereux et nécessitant une réaction immédiate. Briquet (2019) propose ainsi une solution qui permet, une fois l'impact détecté, de classifier cette interaction en utilisant des techniques d'apprentissage supervisé et des réseaux de neurones.

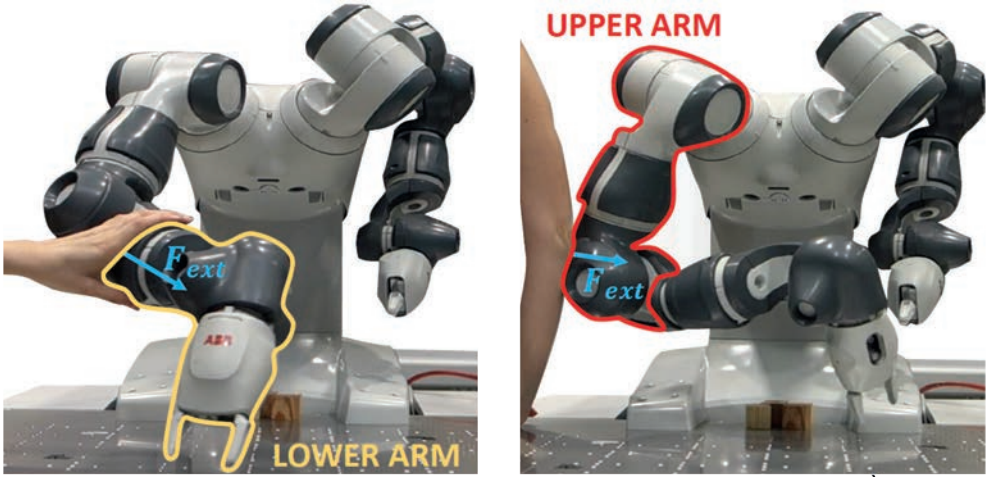


Figure 4 : Utilisation de réseaux de neurones pour la caractérisation du contact. À gauche : une interaction intentionnelle avec le robot. À droite : une collision non désirée (Briquet, 2019).

Au-delà de pouvoir détecter et classifier des contacts avec l'opérateur, le robot doit avoir une connaissance plus fine de son activité afin de rendre l'interaction plus fluide et plus performante. Il s'agit d'analyser les gestes de l'opérateur, d'y répondre correctement, voire de les anticiper. Généralement, des caméras externes sont placées dans l'environnement de travail de l'opérateur

pour détecter ses mouvements. On peut aussi utiliser des capteurs portés par l'opérateur, placés par exemple dans des gants de protection ou sur la tête. C'est le principe mis en œuvre par Coupeté (2016) pour analyser les gestes des opérateurs lors de l'assemblage de pièces automobiles en collaboration avec un robot à deux bras. Des algorithmes basés sur des chaînes de Markov cachées sont utilisés pour classifier les gestes des opérateurs selon des catégories prédéfinies. Les résultats de la reconnaissance servent d'entrées pour le comportement du robot.

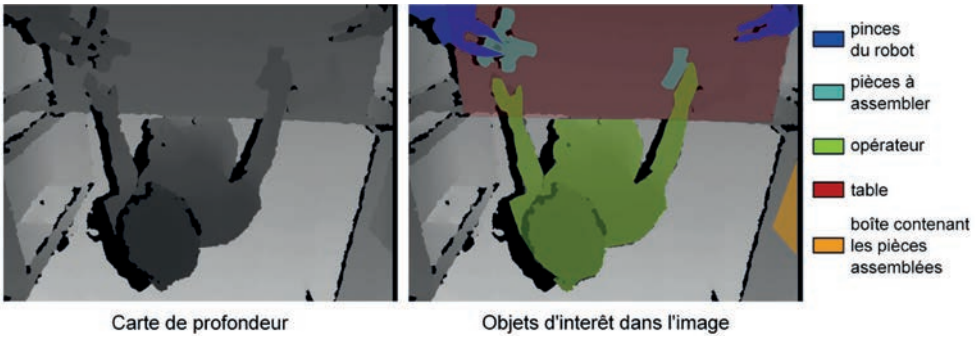


Figure 5 : Utilisation d'une caméra de profondeur pour l'analyse d'activité de l'opérateur et la classification des gestes par chaînes de Markov cachées (Coupeté, 2016).

L'utilisation de l'apprentissage artificiel pour la reconnaissance d'activité n'est pas exclusive à l'interaction homme-robot. Les mêmes techniques peuvent être utilisées pour analyser l'ergonomie du poste de travail, reconnaître l'intention d'une personne à domicile ou encore évaluer la performance d'un sportif ou d'un artiste. Ces techniques de reconnaissance nécessitent néanmoins d'établir des catégories de classification, souvent dépendantes du cas d'étude, ce qui peut pénaliser l'industrialisation de cette méthode.

Pour certains postes de travail, s'adapter aux mouvements de l'opérateur avec qui le robot interagit ne suffit pas : le robot doit également s'adapter au profil et à l'expertise de l'opérateur. En effet, le comportement d'un robot n'a pas le même impact sur l'opérateur selon son niveau d'expertise. Aussi, après plusieurs heures de travail avec le robot, le niveau d'expertise de l'opérateur peut évoluer et le robot doit alors modifier ses paramètres dynamiquement pour s'adapter encore mieux au comportement de l'opérateur. Cette problématique nécessite de pouvoir modéliser finement le profil des utilisateurs. Blanchet *et al.* (2019) proposent une approche de modélisation basée sur l'extraction de données brutes internes au robot, lorsque celui-ci est manipulé par les opérateurs. Des expérimentations ont permis de démontrer l'intérêt de cette approche pour distinguer les opérateurs experts des opérateurs novices, mettant en avant la nécessité d'une assistance différente selon le profil des opérateurs.

Limites et extensions de l'apprentissage artificiel pour la « cobotique »

Un petit nombre d'exemples ont permis d'illustrer l'intérêt des techniques d'apprentissage artificiel pour la « cobotique », que ce soit pour assister à la conception des nouveaux postes collaboratifs ou pour améliorer l'interaction entre l'homme et le robot. Leur description n'est néanmoins pas exhaustive et d'autres applications sont envisageables. Il existe également certaines problématiques pour lesquelles l'apprentissage artificiel peut difficilement être appliqué.

La prise en compte du comportement des opérateurs et de sa variabilité reste aujourd'hui une problématique clé. Les outils actuels de simulation permettent de déterminer de plus en plus

Figure 6 : Utilisation de la réalité virtuelle pour la simulation d'un poste de travail avec un robot collaboratif. L'opérateur interagit ici dans un système immersif à trois faces (Weistroffer, 2014).

fidèlement si un poste de travail collaboratif est acceptable du point de vue de sa sécurité et de son ergonomie, mais ils doivent encore considérer le comportement humain, ses incertitudes et ses erreurs pour donner des résultats plus pertinents. Ces aspects sont aujourd'hui difficilement modélisables dans une simulation. C'est pourquoi la nécessité de mener des tests utilisateurs est toujours présente, soit dans des environnements de réalité virtuelle, soit en situation réelle (Weistroffer, 2014).

Enfin, la sécurité, l'ergonomie et la performance d'un poste collaboratif ne constituent qu'une partie nécessaire au bon déploiement du poste. Des étapes restent encore nécessaires pour que le poste soit considéré comme utilisable et acceptable du point de vue des opérateurs impliqués. Ce sont des critères essentiellement subjectifs et donc difficilement modélisables dans une simulation.

Ainsi, à moins de pouvoir modéliser fidèlement le comportement d'un opérateur, avec ses défauts et ses erreurs, les techniques de simulation et d'apprentissage artificiel devront toujours être accompagnées de tests supplémentaires pour valider ou compléter leurs résultats.

Bibliographie

MAURICE P. (2015), *Virtual Ergonomics for the design of collaborative robots*, thèse de doctorat, Université Pierre et Marie Curie – Paris VI, 208 p.

BRIQUET-KERESTEDJIAN N. (2019), *Impact detection and classification for safe physical Human-Robot Interaction under uncertainties*, thèse de doctorat, Université Paris-Saclay, 247 p.

COUPETE E. (2016), *Reconnaissance de gestes et actions pour la collaboration homme-robot sur chaîne de montage*, thèse de doctorat, PSL Research University, 159 p.

BLANCHET K., KCHIR S., BOUZEGHOUB A., LEBEC O. & HEDE P. (2019), "From Raw Signals to Human Skills Level in Physical Human-Robot Collaboration for Advanced-Manufacturing Applications", *in* GEDEON T., WONG K. & LEE M. (eds.), *Neural Information Processing. ICONIP 2019. Lecture Notes in Computer Science*, vol. 11954, Springer, Cham, pp. 554-565.

WEISTROFFER V. (2014), Étude *des conditions d'acceptabilité de la collaboration homme-robot en utilisant la réalité virtuelle*, thèse de doctorat, École Nationale Supérieure des Mines de Paris, 229 p.

La conduite automatisée, intelligences artificielles et humaines, quelles interactions ?

Par **Antoine LAFAY**
Directeur de la Recherche et de l'Innovation sur le véhicule autonome pour Valeo
et **Guillaume DEVAUCHELLE**
Vice-président Innovation et Développement scientifique du groupe Valeo

Introduction

Avouons-le, se laisser conduire par quelqu'un d'autre n'est déjà pas si facile, alors par une machine... La « voiture autonome » soulève beaucoup d'enthousiasme et autant de contradictions. Le débat est vif et porte sur des plans aussi variés que la faisabilité technologique, l'acceptabilité des usages, l'éthique, le droit, l'impact sur la ville et les infrastructures, les conséquences énergétiques et nous en oublions peut-être.

Cela tient à la place centrale de la mobilité dans nos vies toujours en mouvement. Nous avons tous vécu, chacun à sa façon mais tous intensément, la perte de cette faculté à l'occasion de la Covid-19. Mais en matière de déplacement automobile, la finalité, le déplacement, ne va pas sans le comment, la conduite.

Ne revenons pas sur l'importance de la mobilité et ses multiples dimensions systémiques pour nous concentrer sur la conduite elle-même.

En tant qu'ingénieurs, rationnellement, l'idée même d'un véhicule capable de prendre des décisions extrêmement complexes, en temps réel et aux conséquences potentiellement vitales, nous interpelle. Mais il est clair que l'imaginaire, le subjectif, le plaisir de la conduite, la fascination pour la vitesse, le style, la relation à l'objet dépassent largement la simple rationalité.

Pourtant, le « véhicule autonome » n'est pas nouveau. Le code de la route précise même que le conducteur doit rester maître de son véhicule. Cela vient de la traction hippomobile. Il ne fallait pas que le cheval rentre seul à l'écurie. Implicitement, c'est reconnaître que le cheval a une capacité de décision réelle jusqu'à exécuter seul le trajet. La conduite était alors partagée entre l'homme et l'animal, en constante interaction à l'issue d'un long processus d'apprentissage.

La mécanisation de la traction a pu donner l'impression au conducteur de maîtriser seul la situation, après s'être approprié les réactions d'une mécanique, conçue comme parfaitement prédictible.

Évidemment, cette maîtrise est loin d'être parfaite, et, au fil du temps, un nombre croissant de dispositifs se sont intercalés entre l'homme et la route. Nous les avons oubliés mais les amateurs de voitures anciennes savent ce que c'est que de conduire sans direction ni freinage assistés, sans boîte à vitesses synchronisées...

L'introduction de l'électronique, il y a une trentaine d'années, a changé la donne et a suscité beaucoup de réactions à l'époque. C'est que l'électronique porte une logique propre, le *software,* qui se substitue au conducteur, notamment pour les cas les plus extrêmes de la conduite.

Souvenez-vous des comparatifs entre distances de freinage avec ABS [1] et sans ABS. Avec un champion au volant, le pilote faisait mieux que les premières générations d'ABS, mais tout le monde n'est pas, à chaque instant, un pilote au mieux de sa forme. Les bénéfices sont largement acceptés aujourd'hui et les performances actuelles sont telles qu'il a fallu limiter ou interdire l'utilisation de l'électronique dans la compétition automobile. Bref, chacun confie aujourd'hui à une électronique la « conduite » de son véhicule là où le risque est maximum. Les fonctionnalités sont nombreuses (avec autant d'acronymes : ESP [2], AFU [3]...), et beaucoup sont réglementaires.

L'industrie automobile présente ces applications comme des « aides à la conduite » ou ADAS [4], formule assez pudique, mais la réalité est là : l'électronique gère la direction et le freinage du véhicule dans les cas critiques, en lieu et place du conducteur.

Les régulateurs de vitesse adaptatifs, qui règlent la vitesse du véhicule par rapport au contexte, ont plus récemment soulevé les mêmes inquiétudes mais sont maintenant entrés dans les mœurs. Il ne viendrait plus à l'esprit de personne de qualifier son véhicule d'« autonome » parce qu'il est doté d'un régulateur de vitesse adaptatif et d'un dispositif de maintien dans la file. Nul doute qu'un tel véhicule, s'il avait été disponible il y a 20 ans, aurait été qualifié d'« autonome ».

Il faut donc préciser ce que l'on entend aujourd'hui par « voiture autonome ».

Déploiement des véhicules autonomes

Nous assistons aujourd'hui au déploiement rapide et massif des aides à la conduite pour la voiture de Monsieur Toulemonde, et au développement de quelques véhicules « autonomes ». Le référentiel JSAE [5] définit 5 niveaux d'automatisation de la conduite.

Du niveau 1 au niveau 2, nous parlons d'aide à la conduite : pour le niveau 1, le système assiste le conducteur dans une dimension (contrôle longitudinal) ; pour le niveau 2, le système assiste dans les deux dimensions (contrôle longitudinal et latéral). Pour ces deux niveaux, le conducteur reste toujours responsable, et doit rester vigilant en permanence.

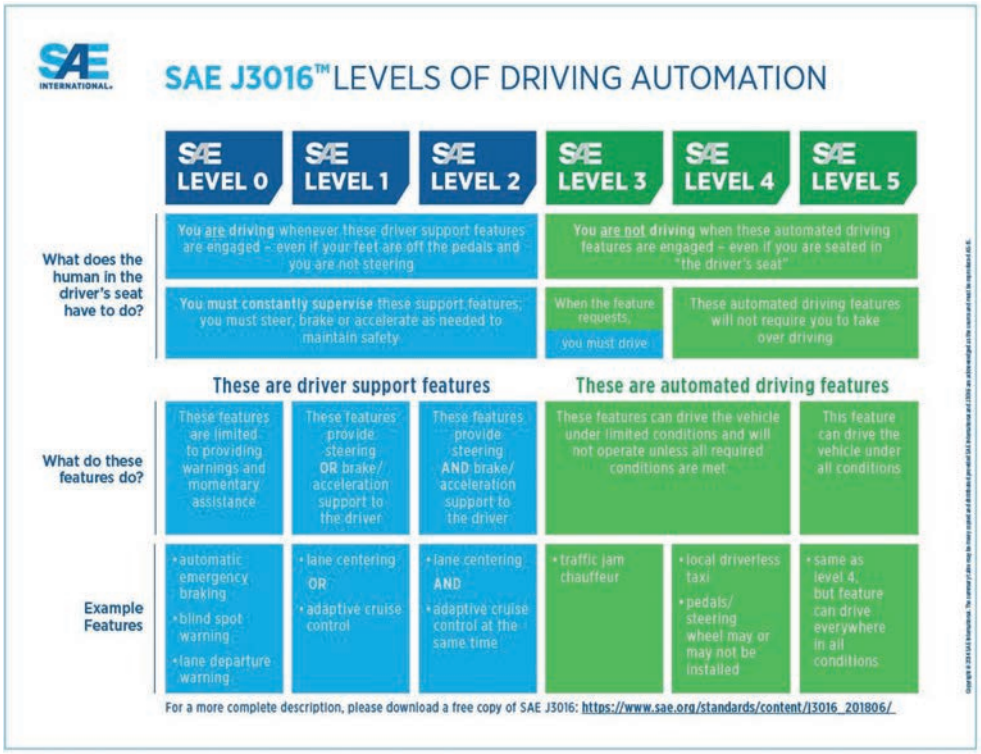
Tout l'enjeu est de garder le conducteur attentif car il doit pouvoir prendre la main en quelques secondes, et de ne pas créer une surévaluation de la performance du système. Les systèmes de niveau 2 ne sont en effet pas assez robustes pour assurer pleinement la sécurité de fonctionnement dans toutes les situations de conduite.

À partir du niveau 3, nous pouvons commencer à parler de véhicule autonome, car le système peut prendre la responsabilité de la conduite et se substituer au conducteur. Dans le niveau 3, le conducteur peut ainsi relâcher sa vigilance temporairement, dans le niveau 4 en permanence, et un véhicule de niveau 5 n'a ni pédale ni volant.

De plus, à partir du niveau 4, les véhicules autonomes feront partie d'une flotte opérée, avec un support de l'infrastructure et d'un opérateur déporté dans un centre de contrôle. L'appellation « véhicule autonome » est donc abusive puisque le véhicule dépend d'un tiers ; « véhicule sans conducteur à bord » serait plus juste.

(1) ABS : Système d'antiblocage des roues. L'acronyme ABS provient de l'allemand *Antiblockiersystem*. Il désigne le système d'anti-blocage des roues permettant d'empêcher, lors d'un freinage pied au plancher, que les roues ne se bloquent.
(2) ESP : *Electronic Stability Program*. En français, il est également appelé « correcteur électronique de trajectoire ».
(3) AFU : Assistance au Freinage d'Urgence
(4) ADAS : *Advanced Driving Assist System* est l'expression anglaise pour désigner les systèmes avancés d'assistance et d'aide à la conduite.
(5) JSAE : *Journal of Society of Automobile Engineers*

L'intelligence artificielle (IA) sera à la fois embarquée dans le véhicule pour toutes les décisions locales, mais également dans l'infrastructure.



https://www.sae.org/

Si les niveaux 1 et 2 sont désormais une réalité, y compris sur des voitures d'entrée de gamme, le niveau 3 vient tout juste de franchir une étape majeure avec l'adoption par les Nations Unies d'une régulation autorisant les véhicules automatisés de niveau 3, équipés de système dit ALKS (*Automated Lane Keeping Systems*). Cela va permettre le déploiement dès 2021 de ces systèmes par tous les grands constructeurs, les premiums en premier lieu.

Des systèmes de niveau 4 sont actuellement en phase de tests, comme des robots-taxis, des navettes autonomes ou des systèmes de parking public avec voituriers automatiques. Ces tests ont lieu soit dans le cadre d'initiatives privées, soit dans le cadre d'expérimentations publiques comme les projets EVRA en France.

Derrière ce concept générique de véhicule autonome, nous pouvons distinguer trois marchés avec des approches différentes :

- Tout d'abord, le véhicule particulier, avec l'approche incrémentale que nous avons vue précédemment, les fonctionnalités sont nombreuses : reconnaissance des panneaux, régulateur de vitesse adaptatif, freinage d'urgence, surveillance d'angle mort, maintien dans la voie, parking automatisé, conduite en embouteillage. Il s'agit ici d'aider, voire de suppléer le conducteur dans certaines des tâches de conduite les plus techniques ou les plus ennuyeuses, et de renforcer la sécurité.
- Ensuite, le transport public, avec une approche beaucoup plus directe. Il s'agit des robots-taxis et des navettes autonomes, fortement automatisés, sans conducteur. Ces véhicules sont acquis

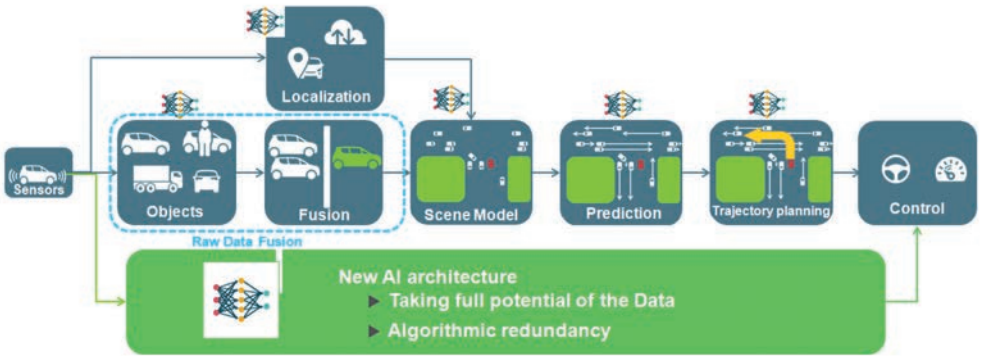
et opérés par des gestionnaires de flotte ou des opérateurs de transport. Sur ces véhicules, les contraintes de coûts et d'intégration sont moins fortes, mais la disponibilité doit être maximale afin de garantir le service et d'assurer la rentabilité du système.

- Enfin, le transport de biens, avec une approche combinant la logistique industrielle et la conduite autonome. Les premiers droïdes de « livraison du dernier kilomètre » sont une évolution des AGV (*Automated Guided Vehicle*) existants dans les usines aujourd'hui. En s'appuyant sur les technologies automobiles, ils effectuent des livraisons en milieu urbain, puis périurbain, avec des missions de plus en plus complexes.

Analyser, comprendre et interagir avec son environnement : la montée en puissance de l'intelligence artificielle

Les véhicules sont équipés d'un nombre de plus en plus important de capteurs, qui ont des capacités de détection au-delà de la perception humaine (caméra thermique, radar, ultrason), générant une quantité extrêmement importante d'informations qu'il faut traiter.

Observons la chaîne algorithmique classique :



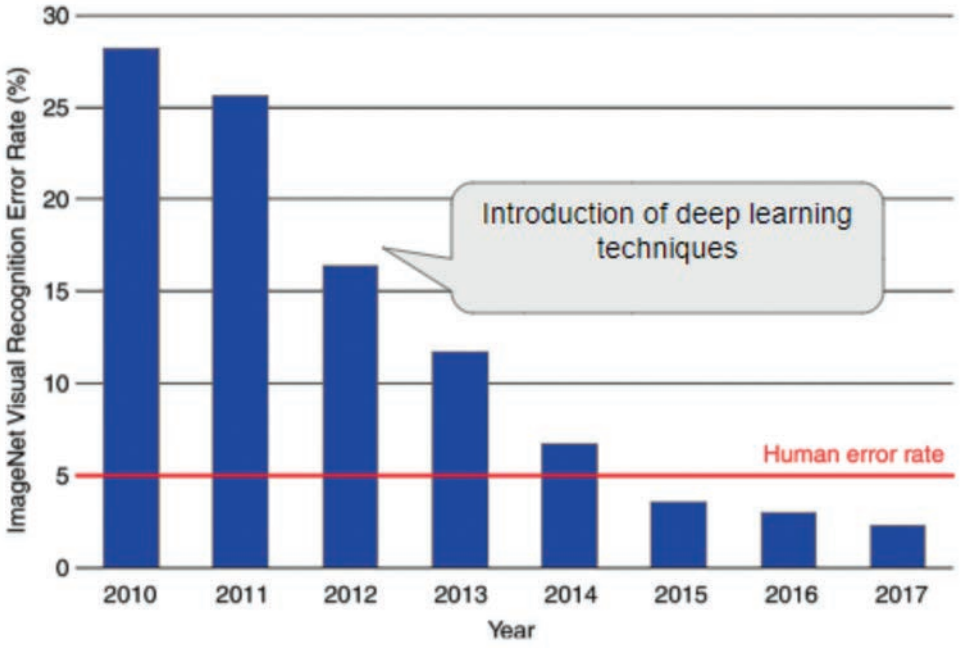
Différentes briques algorithmiques vont être nécessaires :

- Détecter et classifier les objets environnants,
- Localiser,
- Prédire les trajectoires des objets environnants,
- Planifier sa trajectoire,
- Contrôler.

L'essor de l'IA

Pour résoudre ce fabuleux challenge de perception et de prise de décision, l'intelligence artificielle est en train d'être fortement déployée dans l'automatisation des véhicules. Elle devient indispensable pour améliorer la performance et apporter des solutions à des problèmes jusque-là restés sans réponse.

Dans la voiture autonome, elle agit en premier lieu sur la perception. Les performances de détection des algorithmes basés sur du *deep learning* sont grandement améliorées. Sur le graphique ci-dessous, on peut observer l'évolution des performances face au problème de classification d'images de la *dataset* Imagenet. On voit ainsi comment l'introduction des réseaux de neurones profonds a permis d'améliorer très fortement le niveau de classification, jusqu'à des performances surhumaines.



Ce type d'algorithme est aujourd'hui couramment utilisé pour les caméras (frontale et ceinture à 360°) sur les véhicules de série et est en cours de déploiement sur d'autres capteurs de perception comme les Lidars, les radars et les sonars. Le déploiement du *deep learning* dans les autres briques (la fusion, la prédiction, la planification et le contrôle) constitue la deuxième étape.

Enfin, de nouvelles formes d'architecture deviennent possibles grâce aux réseaux de neurones. Par exemple, Valeo a montré lors du CES 2018 que l'on pouvait appliquer une architecture "*end-to-end*", qui consiste à prendre directement l'ensemble des données et d'en sortir des commandes finales. Ainsi, un réseau de neurones prédisait directement le contrôle latéral à partir de l'image en entrée et un autre réseau prédisait le contrôle longitudinal. Le fonctionnement s'apparente aux réflexes humains, issus du cerveau reptilien.

Si l'IA offre de formidables opportunités pour développer de nouvelles fonctions, de nombreux défis restent à résoudre pour exploiter pleinement les bénéfices.

Défi n°1 de l'IA : les algorithmes de *deep learning*

Le *deep learning* est une science récente, en pleine évolution, avec un transfert extrêmement rapide de la recherche à l'industrie. La liaison recherche-industrie est ici essentielle.

Un point fort du *deep learning* est la capacité à adapter les principes et les architectures d'une application à l'autre. Contrairement aux algorithmes classiques qui nécessitent des développements spécifiques à chaque application, les algorithmes de *deep learning* sont transférables d'une application à l'autre.

Ainsi, même si les cas d'usages ne sont pas exactement les mêmes entre l'automobile et l'aéronautique, une grande partie des principes et algorithmes peuvent être communs.

L'adaptation va venir des jeux de données utilisés pour entraîner ces algorithmes. La chaire Drive4All du Laboratoire de Robotique de l'École des Mines a réuni PSA, Safran et Valeo sur ces bases.

Défi n°2 de l'IA : les données

Le carburant de ces algorithmes d'apprentissage est la donnée. Pour entraîner et valider ces algorithmes, il est nécessaire d'avoir un volume de données important et représentatif de l'ensemble des cas que l'on veut traiter.

Des millions de kilomètres sont parcourus pour collecter ces données. La collecte de données n'est rien sans l'annotation de données. Outre la création de l'environnement de collecte de données, il est également nécessaire de créer une solution spécifique pour sélectionner et annoter des centaines de milliers d'images contenant des informations pertinentes.

L'industrie automobile dispose d'un avantage fort qui est la capacité à déployer des systèmes à grande échelle et à récupérer ces données régulièrement, comme le font PSA, Renault et Valeo dans le projet Moove. Et pour encore augmenter la quantité de *data*, les données synthétiques sont utilisées.

Des simulateurs d'environnement multi-physiques permettent de générer une grande diversité de scénarii, en particulier aux limites, dans des cas très rares, extrêmes ou dangereux, qu'il serait difficile d'acquérir en situation réelle.

Défi n°3 de l'IA : la puissance de calcul

L'explosion des volumes de données générées requiert, d'une part, au niveau R&D le développement de solutions de "*data lake*" et de serveurs de calcul pour traiter l'ensemble des données pertinentes, et ainsi d'apprendre les modèles des réseaux de neurones.

D'autre part, pour des systèmes autonomes, il est nécessaire de traiter les données en temps réel. Il faut donc avoir à la fois des solutions algorithmiques d'optimisation et des solutions de calcul *hardware* améliorant la puissance de calcul, la consommation d'énergie et les contraintes d'embarquement.

Plusieurs dizaines de calculateurs se partagent aujourd'hui les différentes fonctions dans un véhicule et proviennent de différents fournisseurs. Les architectures évoluent vers des « contrôleurs de domaine » beaucoup plus puissants, accueillant différentes fonctionnalités.

Défi n°4 de l'IA : l'explicabilité et la validation

Au-delà des performances de ces algorithmes, il faut pouvoir démontrer leur robustesse et leur sûreté de fonctionnement. L'utilisation seule du *deep learning* ne sera pas suffisante.

De nouvelles architectures sont en cours de développement, par exemple le concept de "*Responsibility-Sensitive Safety*" (RSS) de MobilEye.

Les algorithmes de *deep learning* et les algorithmes déterministes plus classiques sont mis en redondance, avec des points de rencontre et de prédiction, permettant l'explicabilité des décisions prises.

Ce thème est repris en France par « le grand défi IA de confiance », dirigé par Julien Chiaroni.

La cybersécurité fait évidemment partie du sujet.

Perspectives

Le champ d'application de l'IA est extrêmement vaste, de la perception à l'intelligence collective, et fait interagir l'intelligence humaine avec l'intelligence artificielle à tous les niveaux.

Pour la perception, comme pour les cinq sens humains, la fusion de données est mieux à même de fournir des informations pertinentes. Les performances de capteurs, comme les caméras, permettent de voir plus loin, plus précisément et de manière beaucoup plus stable que l'œil

humain. D'autres physiques comme l'infrarouge permettent également de beaucoup mieux percevoir l'environnement dans les cas difficiles de brouillard ou de nuit. Est-ce à dire que l'on pourrait conduire « aux instruments » comme le fait l'aéronautique ? Certainement pas souhaitable aujourd'hui !

Il manquerait des éléments essentiels pour la compréhension du contexte, l'analyse des situations, la prévision des trajectoires. L'IA permet maintenant de prédire les intentions d'un piéton par exemple, mais est bien incapable de traiter une foule amassée le long d'un trottoir. Pas facile non plus pour l'IA de détecter tel véhicule avec un comportement agressif, qu'il vaut mieux laisser passer même en ayant la priorité. Le contact visuel reste primordial pour prendre les décisions difficiles.

Au-delà de l'environnement immédiat, pour mieux comprendre le contexte plus global, nous nous sommes déjà habitués depuis longtemps à tenir compte des informations données à la radio, puis du GPS et des applications de navigation. Il ne fait aucun doute que l'IA fournira des informations de plus en plus pertinentes, dont nous ne pourrons plus nous passer. La 5G permettra une forme de téléopération des véhicules.

Au niveau des réflexes, du cerveau reptilien, l'IA *end-to-end* pourrait être beaucoup plus réactive et avec une capacité d'analyse « froide » et non émotive. Qui n'a pas fermé les yeux instinctivement à l'approche d'un danger, se privant ainsi d'informations précieuses ?

L'IA permet maintenant d'analyser la capacité réelle du conducteur à effectuer les tâches de conduite correctement. Les capteurs de vigilance, de stress, entre autres, sont en passe de devenir obligatoires. Nous ne sommes pas toujours les meilleurs juges sur notre propre capacité à tenir un volant. Cela pose évidemment des problèmes d'éthique.

Les possibilités offertes par le *cloud* pour optimiser globalement le transport des personnes et des biens, et offrir les différents services, ne sont limitées que par l'imagination.

Si l'on se place maintenant à un niveau plus philosophique, on ne peut pas réduire la question à la conduite seule. Il ne suffit pas de créer un « cocon » dans un flux de mobilité autour du véhicule et de ses occupants, mais de faire prendre en charge un certain nombre de tâches de conduite par la machine sans que l'homme s'en sente prisonnier ou trop dépendant, et qu'il se sente non seulement en parfaite sécurité, mais aussi à l'aise.

La mobilité a pris une importance telle dans la société moderne que le temps consacré à cette mobilité ne peut pas être un simple intermède entre deux temps « à valeur ajoutée », mais qu'il devienne lui-même un temps à valeur ajoutée.

Le plus simple serait de retrouver dans la voiture automatisée les conditions de tranquillité de la sédentarité « comme à la maison », mais ce serait se priver du plaisir du voyage, du besoin d'évasion, de l'expérience. En quelque sorte, il faut créer un partage des tâches, un hybride où l'homme ne conserverait que le positif du voyage sans les contraintes de la conduite exacerbées par le monde moderne. C'est un beau défi pour l'IA.

La forte progression des marchés ADAS permet à la fois une appropriation progressive des technologies et un financement massif de la recherche dans le domaine de la conduite automatisée et de l'IA. Une révolution silencieuse en quelque sorte.

Mais rappelons-nous qu'émotion vient du vieux français « motion », mouvement, et qu'il n'y a pas de mobilité sans émotion, et pas seulement sur le plan étymologique !

L'intelligence, qu'elle soit humaine ou artificielle, ne fera pas tout.

Stratégie et intelligence artificielle

Par **Henri ISAAC**
PSL, Université Paris-Dauphine

Introduction

En 2017, le fonds d'investissement Deep Knowledge Ventures annonce qu'une intelligence artificielle siège à son conseil d'administration depuis plusieurs années [(1)]. Mark van Rijmenam, fondateur de Datafloq confirme que les intelligences artificielles continueront d'occuper des sièges dans les conseils d'administration [(2)]. Mais paradoxalement, les entreprises qui utilisent le plus d'intelligences artificielles comme Amazon, Facebook, Google, Uber et Apple n'en déploient aucune dans leur conseil d'administration. Ces mêmes entreprises occupent par ailleurs des positions concurrentielles très solides, conquises en peu de temps. Aussi, ce paradoxe illustre bien la difficulté qu'il y a à appréhender le rôle de l'intelligence artificielle et son lien avec la stratégie des entreprises. Les progrès bien réels de l'IA, comme en témoignent GPT-3 d'OpenAI ou encore l'agent de Heron Systems, qui a battu un pilote de l'USAF 5-0 [(3)] au cours d'une simulation de combat aérien, amènent les entreprises à s'interroger sur le rôle de l'IA dans leur stratégie. D'autant que ces méthodes sont désormais largement déployées dans de nombreux secteurs industriels au sein de multiples processus opérationnels (achat publicitaire programmatique, allocation d'actifs financiers, personnalisation des offres en ligne, prise de photo avec *smartphone*, diagnostic médical, fixation des prix, détection de la fraude, etc.) [(4)]. Ces applications nombreuses témoignent donc d'un intérêt réel pour l'intelligence artificielle dans la mise en œuvre de la stratégie de certaines compagnies. Pour autant ces méthodes et technologies modifient-elles la stratégie d'entreprise telle qu'elle a été définie par les sciences de la gestion ? Comment les entreprises peuvent-elles tirer un avantage stratégique de ces méthodes et outils ? Quelles conséquences le développement de telles méthodes peut-il avoir sur le fonctionnement du conseil d'administration et la prise de décision stratégique ? Telles sont les questions que le déploiement de l'IA engendre. Après avoir discuté l'intérêt stratégique de l'IA pour l'entreprise, nous mettrons en évidence l'importance des réorganisations nécessaires pour en tirer un réel avantage compétitif, puis nous montrerons que son usage interroge le plus haut niveau de l'entreprise par les questionnements éthiques que l'IA introduit.

Une nouvelle ère de la stratégie ?

Les incompréhensions nombreuses sur la nature exacte de l'IA, particulièrement répandues dans le monde économique, conduisent souvent à imaginer que l'IA bouleverse les fondements de la stratégie d'entreprise. Si l'IA modifiera indéniablement la logique de certaines industries dans le futur (l'automobile et la logistique d'une façon générale, une partie de la santé et de l'*entertainment*), il n'en demeure pas moins qu'après une décennie de progrès en matière d'IA, la réalité de la transformation de la stratégie des entreprises par l'IA se révèle nettement plus prosaïque. Les

(1) https://asia.nikkei.com/Business/Artificial-intelligence-gets-a-seat-in-the-boardroom
(2) https://www.brinknews.com/will-ai-board-members-run-the-companies-of-the-future/
(3) https://www.youtube.com/watch?v=NzdhIA2S35w
(4) Voir par exemple : "eBay CTO: AI is now an 'ecosystem' for us" https://venturebeat.com/2020/07/17/ebay-cto-ai-is-now-an-ecosystem-for-us/

dernières études en la matière mettent en évidence un usage de l'IA au niveau tactique plutôt qu'au niveau stratégique de l'entreprise, c'est-à-dire essentiellement dans l'amélioration des processus opérationnels de l'entreprise [5].

Décision stratégique et IA

L'IA actuelle produit des modèles prédictifs puissants dans différents domaines (vision, langage, etc.). Les récents progrès pourraient laisser penser que de tels modèles pourraient trouver une application dans la prise de décision stratégique, voire prendre les décisions stratégiques elles-mêmes lorsqu'on les utilise dans un conseil d'administration. Mais ceci nécessite de définir la nature stratégique d'une décision de l'entreprise, ce qui reste encore largement débattu au sein de la recherche en management stratégique depuis de nombreuses années [6]. Dans une perspective où la stratégie est considérée comme la manière de poser les problèmes ou encore la construction d'un récit stratégique dans l'approche du « *sense making* », il apparaît difficile dans de telles perspectives de considérer qu'une IA quelconque puisse jouer un rôle dans la décision stratégique elle-même.

Au-delà de ce que l'on retient comme définition d'une décision stratégique, les méthodes actuelles utilisées dans le développement de l'IA, qu'elles relèvent du *machine learning* ou du *deep learning,* reposent toutes sur des jeux de données nécessaires à l'entrainement en vue de construire un modèle prédictif. Bâtir une IA capable de prendre des décisions stratégiques pour l'entreprise nécessite donc d'avoir accès à un très large spectre de données suffisamment variées et pertinentes pour bâtir un modèle stratégique. Or l'essence même de la prise de décision stratégique est définie par un manque d'informations, une incertitude élevée et une forte interdépendance avec les acteurs de la décision, autant de caractéristiques qui rendent assez improbables de confier les décisions stratégiques d'une entreprise à une IA.

Si un tel choix était effectué, une telle IA ne franchirait pas deux obstacles majeurs pour se substituer aux décideurs : les biais liés aux jeux de données utilisés [7] et la gageure de prédire des phénomènes par nature imprévisibles, comme une pandémie de type Covid-19. La question des biais est certes discutable dans la mesure où les dirigeants d'entreprises possèdent leurs propres biais, cognitifs et décisionnels [8]. On assisterait davantage à une substitution de biais qu'à une suppression des biais.

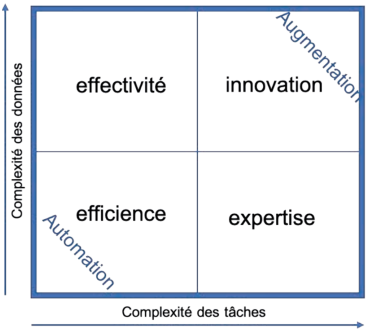
Plus fondamentalement, les ruptures économiques, sociales et politiques qui affectent fortement les décisions stratégiques sont souvent imprévisibles par essence et échappent à toute modélisation aussi puissante soit-elle. Les modèles prédictifs sont limités par le fait qu'ils reposent sur des données du passé dont on sait qu'il ne se reproduit pas (Broussard, 2018). Il suffit pour s'en convaincre d'observer l'échec des modèles de police prédictive, incapables d'appréhender la complexité de la réalité sociale sous-jacente à la criminalité [9].

(5) MIT Technology Review Insights, (2020), *The global AI agenda: Promise, reality, and a future of data sharing*, 20 p., mars.
(6) RUMELT R., SCHENDEL D., TEECE D. (1994), *Fundamental Issues in Strategy: A Research Agenda*, Harvard Business School Press; HAMBRICK D., FREDRICKSON J. (2001), Are you sure you have a strategy? *Academy Management Perspectives*, Vol. 15, n°4, pp.51–62; VAN DEN STEEN E. (2017), "A formal theory of strategy", *Management Science*, Vol. 63, n°8, pp. 2616–2636.
(7) Voir un cas sur la découverte de brevets : CHOUDHURY P., STARR E. & AGARWAL R. (2020), "Machine learning and human capital complementarities: Experimental evidence on bias mitigation", *Strategic Management Journal*, Vol. 41, n°8, pp. 1381-1411.
(8) SCHUMACHER C., KECK S., TANG W. (2020), "Biased interpretation of performance feedback: The role of CEO overconfidence", *Strategic Management Journal*, Vol. 41, n°6, pp. 1139-1165.
(9) BARABAS C. (2020), "Beyond Bias: Re-Imagining the Terms of 'Ethical AI'", *Criminal Law*, http://dx.doi.org/10.2139/ssrn.3377921

Dès lors, l'IA trouvera davantage d'applications dans la définition du contenu des stratégies et leur mise en œuvre que dans le processus de décision stratégique lui-même, même si certains envisagent un pilotage stratégique enrichi par l'IA [10].

Dès lors, il est nécessaire de préciser la nature de ces apports de l'IA à la mise en œuvre de la stratégie, notamment en termes de ressources et de compétences spécifiques modelant certains des choix stratégiques internes et externes.

L'IA comme ressource pour les stratégies



Usages de l'IA en entreprise. Adapté de Accenture 2016

Le développement d'une l'IA repose sur la disponibilité de données existantes, mais aussi sur la complexité de ces données et des tâches qu'elle entend prendre en charge. Si l'IA actuelle excelle dans la réalisation de tâches simples, facilitant leur automatisation, dans le cas des tâches plus complexes, elle ne peut jouer qu'un rôle d'aide ou de complément pour celles-ci, se positionnant alors comme un outil d'augmentation de l'humain. Entre automation et augmentation, quatre types d'usage stratégique sont envisageables.

L'IA joue quatre rôles dans l'exécution des stratégies d'affaires, que l'on présente dans la figure ci-contre.

L'IA comme outil d'efficience

Dans cette perspective, où les données sont peu complexes ainsi que les tâches, l'IA s'envisage comme une nouvelle étape du phénomène d'automatisation grâce au *machine learning*, créant de nouveaux types d'artefacts. De ce point de vue, l'IA constitue souvent une forme de capitalisation de l'expérience de l'entreprise et accélère l'atteinte des effets d'expérience et la baisse du coût moyen de production que la courbe d'expérience traduit. Dès lors, ces nouveaux automates servent idéalement des stratégies de domination par les coûts, accélérant leur baisse sur des opérations comme la logistique et la gestion d'entrepôts, le *merchandising,* la gestion de l'expérience client en ligne. En parallèle de la réduction des coûts, l'IA peut également être le fondement d'une tarification dynamique qui optimise en permanence les prix de l'offre [11]. Dès lors, l'IA s'avère un outil particulièrement pertinent dans les stratégies d'efficience.

L'IA comme outil d'effectivité

Dans cette perspective, l'IA s'envisage comme un moyen de rendre effectif un service et de permettre une expérience fluide de celui-ci pour des millions d'utilisateurs simultanément. À la différence du cas précédent, les données sont plus complexes et souvent non structurées (textes, avis, langage, commentaires). L'automatisation s'appuiera notamment sur les capacités d'apprentissage des habitudes des clients et facilitera ainsi la personnalisation du service. En personnalisant et fluidifiant l'expérience du service, l'IA devient un outil fondamental des stratégies de fidélisation. Les progrès dans les méthodes de reconnaissance du langage laissent présager un usage encore accru de l'IA dans la gestion de la relation client. Une entreprise comme Amazon est illustrative

(10) Voir KIRON D. & SCHRAGE M. (2019), "Strategy For and With AI", *MIT Sloan Management Review*, June, https://sloanreview.mit.edu/article/strategy-for-and-with-ai/

(11) Voir par exemple : LEVINA T., LEVIN Y., McGILL J. & NEDIAK M. (2009), "Dynamic Pricing with Online Learning and Strategic Consumers: An Application of the Aggregating Algorithm", *Operations Research*, Vol. 57, n° 2, pp. 327-341.

d'un tel usage de l'IA, en ce sens qu'elle l'utilise systématiquement pour créer une expérience client la plus fluide possible [12].

De nombreux exemples d'application de ces stratégies d'utilisation de l'IA existent, qu'il s'agisse de la fixation d'un prix pour une location sur Airbnb ou de la personnalisation de l'offre sur Spotify, Amazon ou Uber. Des produits intègrent également des IA pour en améliorer l'expérience. C'est le cas des produits d'Apple qui incluent différentes IA (traitement des images, interface vocale, gestion de l'autonomie de l'appareil) [13] ou encore ceux d'Amazon ou Google.

L'IA comme levier de l'expertise

Dans ce type d'usage, les tâches sont complexes et reposent sur une expertise humaine longue et complexe à produire. Le cas typique est celui du diagnostic médical en radiologie. Il est envisageable d'utiliser l'IA pour faciliter le diagnostic de certaines pathologies. Il n'en demeure pas moins qu'il n'est guère possible actuellement de laisser le diagnostic final à la machine, dans la mesure où la décision affecte directement la personne. Il en va de même pour le conseil financier, où une banque peut encore difficilement laisser la gestion des comptes de ses clients à un automate, aussi puissant soit-il. Dans cette perspective, l'IA complète donc l'action humaine, elle l'augmente.

L'IA comme levier de l'innovation

Utiliser l'IA pour l'innovation est très certainement l'usage qui suscite le plus d'attentes dans certaines industries, comme l'industrie pharmaceutique, où il existe plus de 200 *start-up* spécialisées dans la découverte de nouvelles molécules thérapeutiques [14]. Dans ce cas, l'IA est capable de combiner des données qui dépassent ce qu'un humain peut traiter et d'ouvrir la voie à de nouveaux traitements [15]. Dans la conception des objets, l'IA est aussi potentiellement utilisable comme l'exemple du développement d'une structure robuste et très original de véhicules par une IA le laisse envisager [16]. De tels usages pourraient potentiellement se révéler disruptifs, dans la mesure où ils modifieraient profondément les processus de création de valeur.

L'accès à de nouvelles ressources et compétences modifie les stratégies d'entreprise

Les quatre scenarii d'usage de l'IA sont précisément ceux que déploient les plateformes numériques dans différentes industries (voyages, hôtellerie, commerce, transports, etc.). Conçues à l'ère numérique, celles-ci ont bâti des systèmes d'information aptes à collecter les données nécessaires aux modèles prédictifs qui nourrissent l'expérience des clients. Dès lors, ces plateformes possèdent un avantage concurrentiel face aux entreprises historiques en collectant massivement des données, car leur modèle opérationnel est en partie conçu pour cette finalité [17]. Ces acteurs ont transformé la donnée en une véritable ressource stratégique, modifiant ainsi le jeu concurrentiel.

(12) Voir LEVY S. (2018), "Inside Amazon's Artificial Intelligence Flywheel. How deep learning came to power Alexa, Amazon Web Services, and nearly every other division of the company", *Wired*, https://www.wired.com/story/amazon-artificial-intelligence-flywheel/

(13) AXON S. (2020), "Here's why Apple believes it's an AI leader—and why it says critics have it all wrong", https://arstechnica.com/gadgets/2020/08/apple-explains-how-it-uses-machine-learning-across-ios-and-soon-macos/

(14) https://blog.benchsci.com/startups-using-artificial-intelligence-in-drug-discovery

(15) HEILWEIL R. (2020), "Would you take a drug discovered by artificial intelligence? An OCD drug created via AI will be tested on humans", https://www.vox.com/2020/1/31/21117102/artificial-intelligence-drug-discovery-exscientia

(16) HERRERA A. (2019), "A Compelling Example of AI in CAD: Autodesk's Take on Generative Design", https://www.cadalyst.com/digital-design-solutions/product-design/compelling-example-ai-cad-autodesk%25E2%2580%2599s-take-generative-design-?page_id=3

(17) ISAAC H. (2021), *Les business models des plateformes*, Vuibert, à paraître.

Sans données, il est difficile pour les concurrents de bâtir une IA quelconque. L'accès aux données peut alors parfois s'interpréter comme une barrière à l'entrée [18].

Dans cette perspective, les acteurs historiques font souvent face à un manque de données pour rivaliser avec ces nouveaux joueurs. Dès lors, ils sont amenés à envisager des stratégies de partage de données pour compléter leurs jeux de données et bâtir des modèles prédictifs fiables. La constitution de « *pools* de données » a d'ores et déjà débuté dans certaines industries comme la publicité, où des alliances entre différents acteurs construisent des jeux de données élargis pour pouvoir rivaliser avec les nouveaux acteurs comme Facebook et Google [19]. Si ces alliances peuvent parfois constituer des difficultés du point de vue du droit de la concurrence, elles sont plutôt vues en Europe comme un moyen de rivaliser avec les plateformes américaines et chinoises [20] et constituent un élément central de la feuille de route numérique de la Commission européenne [21].

Au-delà de la collecte des données, c'est la capacité à modéliser celles-ci qui forme le fondement d'une IA. Dans cette perspective, force est de constater que les plateformes numériques ont largement identifié ces besoins et recruté de nombreux experts. Plus encore, elles ont acquis de nombreuses entreprises spécialisées afin d'accélérer le développement de leurs savoir-faire en la matière, tant ces savoir-faire sont longs, complexes et coûteux à développer en interne [22]. Ces politiques d'acquisition sont parfois perçues comme des politiques concurrentielles prédatrices dans le sens où elles réduiraient l'intensité concurrentielle [23], mais d'aucuns considèrent qu'il s'agit d'une réponse aux défaillances du marché de la formation et du travail, ne produisant pas suffisamment de compétences nécessaires aux besoins des entreprises.

Si l'arrivée à maturité des méthodes d'IA a déplacé une partie du jeu concurrentiel, ce qui semble désormais décisif dans la réussite de l'IA dans une perspective stratégique, c'est la capacité de l'entreprise à intégrer et à développer les compétences nécessaires à la création et à l'exploitation d'une IA pour en faire une compétence distinctive sur le marché.

Réussir l'intégration de l'intelligence artificielle dans la stratégie d'entreprise

L'utilisation actuelle de l'IA dans les entreprises est surtout le fait d'entreprises numériques récentes construites autour des technologies de l'information. Dès lors, la question de l'organisation ne se pose pas dans les mêmes termes que pour les entreprises traditionnelles tant du point de vue organisationnel que technique.

Du point de vue organisationnel (Kretschmer et Khashabi, 2020), déployer les techniques de l'IA nécessite de rassembler des compétences et des ressources spécifiques (Tambe, 2014 ; Rock, 2019). En effet, les équipes IT ne sont pas aptes à elles seules à déployer de telles techniques, car

(18) MANKE R. (2015), "Big Data as a Barrier to Entry", Competition Policy International, September, https://www.competitionpolicyinternational.com/big-data-as-a-barrier-to-entry/ - voir également CASADO M. & LAUTEN P. (2019), "The Empty Promise of Data Moats" : https://a16z.com/2019/05/09/data-network-effects-moats/

(19) Garivity en France, Skimlinks au Royaume Uni

(20) VESTAGER M. (2016), "Big Data and Competition", https://ec.europa.eu/commission/commissioners/2014-2019/vestager/announcements/big-data-and-competition_en

(21) Commission européenne (2020), « Une stratégie européenne pour les données », février, 42 p.

(22) CROCHET-DAMAIS A., (2020), « Apple, le Gafam ayant acquis le plus de start-up dans l'IA depuis 2010 », https://www.journaldunet.com/solutions/dsi/1492969-apple-le-gafam-ayant-acquis-le-plus-de-start-up-dans-l-ia-selon-satista/

(23) BOURREAU M. & DE STEEL A. (2020), "Big Tech Acquisitions: Competition and Innovation Effects & EU Merger Control", https://www.cerre.eu/publications/big-tech-acquisitions-competition-and-innovation-effects-eu-merger-control; Digital Competition Expert Panel, (2019), Unlocking digital competition, www.gov.uk/government/publications; ARGENTESI E., BUCCIROSSI P., CALVANO E., DUSO T., MARRAZZO A., NAVA S. (2019), "Ex-post assessment of merger control decisions in digital markets", June, https://www.learlab.com/publication/ex-post-assessment-of-merger-control-decisions-in-digital-markets/

celles-ci reposent sur une combinaison de trois types de compétences : mathématiques/statistiques, informatique et métiers. La plupart des entreprises confient cette responsabilité à une équipe *data science* regroupant ces compétences sous l'autorité d'un *chief data officer* (CDO).

Cependant, la création d'une nouvelle fonction ne produit pas *per se* de l'IA. Deux écueils récurrents existent. L'absence de données et/ou la médiocrité de la qualité des données internes disponibles représentent un premier écueil très répandu, nécessitant des refontes du système d'information et de l'organisation, mais plus encore une culture de la donnée à tous les niveaux de l'entreprise [24]. Le second écueil est le passage d'un modèle prédictif réalisé dans un *data lab* à un modèle intégré dans les processus opérationnels de l'entreprise, et donc dans son système d'information. Cette phase d'industrialisation de l'IA est une phase cruciale dans le déploiement de l'IA et repose en grande partie sur des compétences informatiques mais aussi de conduite du changement, tant les conséquences de son introduction peuvent modifier l'exécution des processus et la relation client.

D'un point de vue technique, le déploiement de l'intelligence artificielle repose à ce jour sur trois alternatives principales :

- Utiliser une solution prête à l'emploi proposée par des éditeurs spécialisés comme la détection de fraude dans l'assurance ou l'optimisation des enchères dans la publicité en ligne ;
- S'appuyer sur des offres sur étagère proposées par les géants du numérique (Amazon, Microsoft, IBM, Alibaba) ou des *start-up* spécialisées, reposant sur des modèles de traitement préentrainés qu'il faut affiner avec des données internes ;
- Construire sa propre solution à partir des briques en libre-service pour créer ses propres modèles tout en exploitant les architectures techniques des acteurs du *cloud* (Amazon, Microsoft, Google).

Les conséquences stratégiques de ce choix entre ces trois alternatives sont importantes. En effet, dans le premier cas, l'entreprise possèdera les mêmes outils que ses concurrents et ne pourra pas bâtir une offre différenciante sur cette base. Elle ne pourra être qu'au standard du marché. Dans le second cas, la qualité des données internes est fondamentale pour bâtir des modèles prédictifs fiables, ce qui nécessite une parfaite gestion de la chaîne de valeur de la donnée. Dans le troisième cas, le rôle des compétences spécialisées se révèle crucial dans la capacité à créer des modèles spécifiques à l'entreprise.

Gouvernance de l'entreprise et IA

Le développement rapide des méthodes d'IA interroge les choix stratégiques des entreprises, mais aussi les méthodes de prise de décision et leur déploiement au sein de l'organisation. Les organes de gouvernance, au premier rang desquels le comité de direction et le conseil d'administration, sont par conséquent concernés par ces évolutions. Sans aller jusqu'à introduire une IA en leur sein, les enjeux sont nombreux. Le premier enjeu est un enjeu de compétences des dirigeants sur ce sujet très technique, où la technologie emporte souvent l'essentiel de la discussion alors même que les enjeux se situent souvent sur les données nécessaires aux IA et sur l'organisation et la gestion des compétences des équipes en charge de ces méthodes. Comment la gouvernance maintient-elle ses connaissances sur un sujet de plus en plus stratégique, qui requiert des expertises qui dépassent largement celles de ses membres ?

Les entreprises ont ces dernières années tenté l'introduction d'une fonction centrée sur la numérisation (*chief digital officer*) ou sur les données (*chief data officer*). Si les *chief digital officers*

(24) Voir STRENGOLTH P. (2020), *Data Management at scale*, O'Reilly, 300 p.

ont plutôt tendance à s'estomper au profit des *chief data officers,* il ne faut pas enfermer ceux-ci dans un nouveau rôle de DSI bis, mais bien dans celui d'un architecte des données au service de la création de valeur et donc de l'IA. Il apparaît ainsi nécessaire que le *chief data officer* possède une solide culture en matière d'IA afin d'éclairer les discussions de la gouvernance sur ce sujet stratégique.

Faut-il utiliser l'IA ? Comment rivaliser avec des IA concurrentes ? Faut-il utiliser des IA disponibles sur le marché ? Comment les utiliser dans le cas spécifique de l'entreprise ? Jusqu'où les utiliser ? Telles sont les questions qui se posent à la gouvernance. Ces questions en ouvrent rapidement d'autres plus spécifiques sur les critères de la décision. La nécessité de principes éthiques liés à l'utilisation de la donnée réside dans le fait que les modèles prédictifs et l'intelligence artificielle issus des données génèrent potentiellement de nombreux problèmes, qui affectent directement les utilisateurs de ces systèmes et les décisions les concernant [25]. Par exemple, faut-il, pour une banque, bâtir une offre de service sur le fait de pouvoir prédire de façon fiable un divorce ? Jusqu'où aller dans le pilotage d'un compte client dont une IA est capable de prédire le solde mensuel ? Ces questions dépassent la seule question de la faisabilité et de l'acceptabilité ; elles ouvrent des questionnements éthiques, sur le sens de l'action d'une entreprise et le libre arbitre du client.

Définir un cadre éthique est donc une nécessité absolue pour les dirigeants s'ils veulent pouvoir prendre des décisions en matière d'IA, dont ils pourront rendre compte devant leurs parties prenantes et en premier lieu les clients et les employés.

Dans cette perspective, la mise en œuvre d'un comité d'éthique de la donnée en charge de la définition des principes et de leur déclinaison dans les différentes politiques, au sein des processus métiers et des formations, est une nécessité.

Plusieurs principes éthiques se dégagent des différents travaux d'instances variées sur le sujet [26]. Ils sont résumés dans le tableau suivant :

Transparence	Explicabilité des modèles utilisés. Information sur l'utilisation d'un algorithme dans la prise de décision concernant une personne
Équité	Gestion des biais liés aux données afin d'éviter les risques de discriminations
Non-malfaisance	Principes d'Asimov. Aucun usage de la donnée ne doit mettre en danger un humain ou diminuer l'intégrité physique et mentale
Responsabilité	Principe de redevabilité
Liberté et autonomie	Principe du consentement et d'auto-détermination
Confiance	Principe de confiance
Dignité	Respect de la dignité humaine

Tableau 1 : Principaux principes éthiques dans l'utilisation des données

De tels principes doivent s'intégrer dans les méthodes de conception et de production des algorithmes utilisant massivement des données (*privacy by design*, *security by design*, *inclusiveness by design)*, et faire l'objet d'une évaluation préalable (*Privacy Impact Assessment* [27], *Equality Impact Assessment* [28]). Il importe que l'éthique et la conformité ne se positionnent pas comme des censeurs permanents des équipes de développement, mais plutôt comme des accompagnateurs

(25) WHITTLESTONE J., NYRUP R., ALEXANDROVA A., DIHAL K. & CAVE S. (2019), *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research,* London: Nuffield Foundation.
(26) Voir une synthèse : JOBIN A., IENCA M. & VAYENA E. (2019), "The global landscape of AI ethics guidelines", *Nature Machine Intelligence*, https://doi.org/10.1038/s42256-019-0088-2
(27) https://www.cnil.fr/fr/RGPD-analyse-impact-protection-des-donnees-aipd
(28) https://en.wikipedia.org/wiki/Equality_impact_assessment

permettant de développer l'IA de façon pertinente et conforme, en respectant les principes éthiques que l'organisation aura choisi de suivre et dont elle pourra rendre compte à ses parties prenantes.

Ainsi, l'IA actuelle fournit le plus souvent des avantages d'ordre tactique plutôt que stratégiques. Le déploiement de telles technologies repose sur de larges jeux de données, véritables ressources stratégiques, souvent mal appréhendées par les entreprises. Dès lors, d'importantes réorganisations se révèlent essentielles pour tirer avantage de l'IA. Les ressources et les processus nécessaires à son élaboration et à sa gestion requièrent en effet de profondes modifications dans la gestion de la chaîne de valeur des données. Ces transformations sont tout autant des changements techniques que culturels, ce qui explique en partie les difficultés à construire un avantage compétitif pour nombre d'entreprises. Enfin, lorsque les ressources et compétences ainsi que les processus sont réunis, il importe de comprendre les limites de ces technologies et de construire une doctrine d'emploi qui respecte un certain nombre de valeurs, afin que son acceptabilité soit la plus aisée possible.

Bibliographie

BATALLER C. & HARRIS J. (2016), "Turning Artificial Intelligence into Business Value. Today", Accenture.

TAMBE P. (2014), "Big Data Investment, Skills, and Firm Value," *Management Science*, Vol. 60(6), pp. 1452-1469.

PERRAULT R., SHOHAM Y., BRYNJOLFSSON E. *et al.* (2019), "Artificial Intelligence Index 2019 Annual Report," Human-Centered Artificial Intelligence Institute, Stanford, California: Stanford University, December.

ROCK D. (2019), "Engineering Value: The Returns to Technological Talent and Investments in Artificial Intelligence," unpublished working paper, MIT Sloan School of Management, Cambridge, Massachusetts, May.

BROUSSARD M. (2018), *Artificial Unintelligence: How Computers Misunderstand the World,* MIT Press.

KRETSCHMER T. & KHASHABI P. (2020), "Digital Transformation and Organization Design. An Integrated Approach", *California Management Review*, http://dx.doi.org/10.2139/ssrn.3437334

A Note on the Interpretability of Machine Learning Algorithms

By **Dominique GUÉGAN**
University Paris 1 Panthéon-Sorbonne and University Ca'Foscari in Venezia

Introduction

Following OECD (2016) "Artificial Intelligence (AI) is the ability of machines and systems to gain and apply knowledge and to carry out intelligent behavior". Applications range from education to social welfare, to energy and the environment. Advanced developments in the field of Machine Learning (ML) imply that machines will be able to learn from their experience and make their own decisions, without further input from humans, beyond initial design of the machine. Already, machines have surpassed the ability of humans to perform certain functions, such as image recognition and other intelligence-related tasks.

At the same time, generation of large volumes of data and the creation of centralized data repositories promise to drive growth across all sectors of society including agriculture, industry, banking, resource allocation, public health, education, and poverty reduction. Specifically, these data can be used to determine relationships, predict behaviors and outcomes, and establish dependencies between correlated variables. Algorithms like those developed through ML are used to generate automated outcomes using these data and improve the performance of algorithm-driven tasks, promoting improved business operations, management, and productivity as well as improved consumer-driven tasks.

Despite the growth of the benefits engendered by AI and the development of ML, some fears appear in the society revealed in particular by the press. Indeed, looking at it we observe that the influence and the importance of AI appears through disturbing titles like "software uses across the country to predict future criminals, is biased against black people", or "if you are not a white male, AI's use in healthcare could be dangerous for you", or "algorithms are making the same mistakes assessing credit scores that humans did a century ago", etc. Thus, even if AI creates advantages in the day life, it questions several issues, in particular a main question is: are the predictions provided by AI biased? A common idea is that the softwares are supposed to make policy more fair and accountable, but a huge literature shows that the predictions can be unfair, and coming from society expectations are very high, (Hardt *et al.,* 2016), (Zafar *et al.,* 2017), (Agarwal *et al.,* 2018) or Berg *et al.* (2018) with their works on discrimination, and also Miller (2019) among others.

In many cases, the reasons evocated to use AI consider that they can allocate resources with higher precision, can reduce the role of human instincts and prejudices, but in counterpart perpetuate biases against certain groups (for instance in case of racial profling). The opportunities associated to the use of AI concern higher accuracy, effectiveness, lower cost, higher effciency, preventing human biases and prejudices, transparency, consistency, more equal access to opportunities and resources. But at the same time we observe unfairness, unequal allocation of benefit or harm, opaqueness, inexplicability, unequal representation, invasion of privacy. Thus, it is important to focus on technical solutions to enhance fairness, explainability and accountability for ML systems, because if technical tools are useful, they are not sufficient.

Another reason of the necessity to understand how AI works refers to the potential impact that the European Union's new General Data Protection Regulation will have on the routine use of machine learning algorithms. It takes effect as law across the EU in May 2018 and has, in particular, the objective to restrict automated individual decision-making (that is, algorithms that make decisions based predictors user-level) which can "significantly affect" users. The law has created a "right to explanation" whereby a user can ask for an explanation of an algorithmic decision that was made about them. Interesting discussions on this subject can be found in Goodman & Flaxman (2017) and Whatcher *et al.* (2017) for instance.

Thus, three important topics emerge for acceptation of the use of ML by the scientist, industrial and regulatory communities. They concern issues of fairness, explainability, and accountability. The sources of unfairness are (i) data unfairness, (ii) algorithmic unfairness, (iii) impact unfairness. In this paper we are interesting by the question of interpretability. Indeed the concept of opacity seems to be at the very heart of new concerns about 'algorithms' among legal scholars, social scientists and engineers. Using the data as inputs, the algorithms produce an output (a classification i.e. whether to give an applicant a loan, or whether to tag an email as spam) or predictions. The output of the algorithm rarely does one have any concrete sense of **'how'** or **'why'** a particular decision has been arrived at from inputs. Additionally, the inputs themselves may be entirely unknown or known only partially. The question naturally arises, what are the reasons for this state of not knowing? Following recent researches we discuss some solutions related to these questions.

In order to provide global insight on the subject we begin to recall some definitions of fairness and provide some references on the subject in Section two. Section three is devoted to the definition of interpretability. In Section four we propose solutions for the interpretability of the algorithms. Section five focuses on two local solutions whose interest could be determinant for regulations. Section six gives an idea of the future possible ways to measure the interpretability of the algorithms. Section seven concludes.

Unfairness

Applications of fair machine learning, in the literature, concerns recidivism prediction, automated hiring, and face recognition (among others), where fairness can be understood, at least partially, in terms of well-defined quantitative metrics. However it has recently been shown that algorithms trained with biased data have resulted in algorithmic discrimination, in particular the statistical methods used in the US judicial system, pointing to the bias against African-American accused, considering that African-American accused were more likely to be wrongly labeled as higher risk of recidivism (Wadsworth *et al.,* 2018).

Thus, significant effort in the fair machine learning community has focused on the development of statistical definitions of fairness (Hardt *et al.,* 2016; Berk *et al.,* 2018) and algorithmic methods (Agarwal, 2018; Kusner *et al.,* 2017). The first notion of fairness which was introduced is "statistical parity", called also "group fairness" or "demographic parity" which equalizes outcomes across protected and non protected groups. Demographic parity requires that a decision is independent of a protected attribute meaning that membership in a protected class should have no correlation with the decision. Nevertheless this approach can create highly undesirable decision: for instance if the protected attribute is gender, one might incarcerate women who pose no public safety risk so that the same proportions of men and women are released on probation (Dwork *et al.,* 2012).

In order to avoid the limitation of the previous definition the "equalized odds" with respect to a protected attribute was introduced: the predictor and the protected attribute are independent conditionally on the output (Hardt *et al.,* 2016). An unfairness metric, which is defined in terms of

misclassification rates, has been introduced by Zafaz *et al.* (2017) called "disparate mistreatment". The authors call a decision-making process to be suffering from disparate mistreatment with respect to a given sensitive attribute (*e.g.,* race) if the mis-classification rates differ for groups of people having different values of that sensitive attribute (*e.g.,* black or white). Fermanian and Guégan (2020) provide updates on the subject.

To avoid unfairness in the ML process two stategies have been developed. The pre-process training ensures fairness of any learned model eliminating any sources of unfairness in the data before the algorithm is formulated. A major problem with this approach is that interaction effects (*e.g.,* with race and gender) containing information leading to unfairness are not removed unless they are explicitly included in the residualizing regression even if all of the additive contaminants are removed (Zemel *et al.,* 2013; Berk *et al.,* 2018; Lu *et al.,* 2016). In the post-processing training, after the algorithm is applied its performance is adjusted for instance by random reassignment of the class label previously assigned by the algorithm to make it fair (Feldman *et al.,* 2015; Hardt *et al.,* 2016). For instance a decision tree learner can be changed in splitting its criterion and pruning its strategy by using a novel leaf re-labeling approach after training in order to satisfy fairness constraints, (Kamiran *et al.,* 2010; Zliobaite, 2015; Agarwal *et al.,* 2019; Chzhen *et al.,* 2020).

However, it has been largely documented that simply removing certain variables from a model does not ensure predictions that are, in effect, uncorrelated to those variables. For example, if a certain geographic region has a high number of low income or minority residents, an algorithm that employs geographic data to determine loan eligibility is likely to produce results that are, in effect, informed by race and income (Hardt *et al.,* 2016). As unfairness is also a part of the non interpretability of the algorithms we question now the notion of interpretability and introduce some solutions.

A tentative to define interpretability

The need for explaining the decisions of expert systems was discussed as early as the 1970's. Nevertheless the definition associated to interpretability is not fixed. Following Biran and Cotton (2017), we can summarize most of the discussions around this concept as: "A key component of an artificially intelligent system is the ability to explain the decisions, recommendations, predictions or actions made by it and the process through which they are made." This ability can concern (i) the interpretability associated to justification which explains **'why'** one obtains such result, or (ii) the interpretability associated to explanation which corresponds to **'how'** we got this result. In that latter case, we focus on post hoc explanations. Interpretability can also refer to transparency which is the opposite of opacity or 'black-boxness', including the knowledge of the entire model, the knowledge of individual components such as parameters, and the knowledge of the training algorithm (Burrell, 2016; Doshi-Velez and Kim, 2017; Lipton, 2018).

Going through the literature, we see that there is no consensus on the definition of interpretability. It seems reasonnable to consider on the one hand the **explicability** which specifies **how** the model works, on the other hand **understanding** of the predictions **why**, which is a matter of **interpretation.** Discussing these two issues is important; indeed, in some application domains users need to understand the system's recommendations enough to legally explain the reason for the decisions. For instance in the medical domain, if a doctor makes a decision (say, recommends surgery) based on the prediction of a classification model and that leads to major harm to the patient, the doctor should understand the reason for the model's predictions in order to defend her/his decisions in court if she/he is sued for medical negligence. Legal requirements are also common in credit scoring applications, where a bank often has the legal obligation of explaining why a customer was denied credit.

Nevertheless even if the GDPR recitals state that a data subject has the right to "an explanation of the decision reached after algorithmic assessment," this requirement prompts the question: what does it mean, and what is required to explain an algorithm's decision? In fact a legally binding right to explanation does not exist in the GDPR, and we can consider that the right would only apply in limited cases: when, for instance, a negative decision was solely automated and had legal or other similar significant effects. Thus, explaining the functionality of complex algorithmic decision-making systems and their rationale in specific cases is important even if it is a technically challenging problem. A black box predictor is a data-mining and machine-learning obscure model, whose internals are either unknown to the observer or they are known but uninterpretable by humans. An explanation has to be an "interface" between machines and a decision maker that is comprehensible to humans (Whatcher, 2017).

In the state of the art a small set of models are considered easily understandable and interpretable for humans: decision tree, rules, linear models. A decision system based on a decision tree exploits a graph structured like a tree and composed of internal nodes representing tests on features or attributes (*e.g.,* whether a variable has a value lower than, equals to or greater than a threshold) and leaf nodes representing a class label. A decision tree can be linearized into a set of decision rules with the if-then form for instance.

Another set of approaches adopted to provide explanations are linear models. This can be done by considering and visualizing the features importance, *i.e.,* both the sign and the magnitude of the contribution of the attributes for a given prediction. If the contribution of an attribute-value is positive, then it contributes by increasing the model's output. Instead, if the sign is negative then the attribute-value decreases the output of the model. An intrinsic problem that linear models have when used for explanation is that when the model does not optimally fit the training data, it may use spurious features to optimize the error, and these features may be very hard to interpret for a human. Nevertheless these models can be used to interpret more complex models like Support Vector Machine, Deep Neural Networks or Convolution Neural Networks for example. We can distinguish two approaches the global or local approaches to make the algorithms more interpretable.

Does it exist interpretable solutions?

When a model is completely interpretable we are able to understand the whole logic of a model and follow the entire reasoning leading to all the different possible outcomes. In this case, we are speaking about global interpretability. At the contrary the local interpretability corresponds to the situation in which it is possible to understand only the reasons for a specific decision: only the single prediction/decision is interpretable. In the first case we face the "black box explanation problem" which consists in providing a global explanation of the black box model through an interpretable and transparent model: it **explains the model** (**how** it works). The second case concerns "the outcome explanation problem" which consists in providing an **explanation for the outcome** of the black box (**Why** this result). In the former case the solutions mimic the black box, and in the latter case the solutions provide a predictor which is locally interpretable.

We precise now the formal framework used in the rest of the paper. Let be X the vector of original inputs, $X \in \boldsymbol{X} \subset R^d$, we denote $X' \in \boldsymbol{X}$ a vector for its interpretable representation. The classifier assimilated to a blackbox need to be explained is designated by $b : \boldsymbol{X} \rightarrow R$ and $Y = b(X)$, shorthand for $Y = \{b(x) \mid x \in \boldsymbol{X}\}$, $b(X)$ is a black box predictor, whose internals are either unknown to the observer or they are known but uninterpretable by humans. Y is commonly called the output. The objective is to find an interpretable classifier $c : \boldsymbol{X} \rightarrow R$, $c \in C$, where C is a class of potentially interpretable models such as linear models, decision trees, or rule lists, and let Ω(C) be a measure

of the complexity (as opposed to interpretability) of c (deph of a tree, number of non-zeros in a linear model, etc). Thus c is an interpretable predictor yielding a decision c(X) = Y which can be given a symbolic interpretation comprehensible by a human, *i.e.*, for which a global or a local explanation is available.

Working with supervised ML, one uses a trained data set to train the predictor b, and a test data set D^T to evaluate its performance : let $\hat{Y}$ the outcome using the training data set, and $(X,\hat{Y}) \in D^T$. The objective is to verify how matches $\hat{Y}$ and Y=b(X). The interpretable predictor c need to be as close as possible to b in the sense that c(X) = b(X), for $(X,\hat{Y}) \in D^T$ and to mimic the results obtained through b.

A **global** interpretable model c_g is such that c_g = f (b,X), for some strategy f(.,.), and is derived from b and a subset of $X \in \boldsymbol{X}$. A **local** interpretable predictor is defined as c_l = f(b, X) derived using b and a **neighborhood** of x. The objective is to identify the function f. For simplicity, an input will be designated by x in the following. We introduce several strategies f permitting to determine global c_g or local c_l solutions and then propose specific solutions for interpretability or explainability.

Global approach

For instance, explaining neural networks with decision trees is a global approach. In that case, to build f a sort of "prototype" is generated for each target class in Y by using genetic programming to query the trained neural network b; the input variables X are exploited for constraining the prototypes; then the best prototypes are selecting for inducing the learning of the decision tree c_g. This approach leads to get more understandable and smaller decision trees starting from smaller data sets. Since 1996 single tree approximations for NN have been developed, Craven *et al.* (1996). New approaches are detailed in Guidotti et al. (2018).

Other approaches use decision rules. The solution f lies on a fair documentation of the process by inserting knowledge into neural networks, extracting rules from trained NNs, and using them to refine existing rules. Same kind of processes using rules-based classifiers have also been developed for tree ensemble or Support Vector Machines. All these solutions are not generalizable because they are strongly dependent on the black box b and on the specific type of decision rules c_g. This is a limitation of the method, thus **agnostic** methods have been developed with the objective to be ajustable for **all models**. The Generalized Additive Models (GAM) approach proposes a global solution based on splines functions, as bagged and boosted ensembles trees that chooses the number of leaves adaptively. If we denote c_j(x) the output obtained from tree j, j = 2, … , M, for classification purpose, the output b of the additive tree model is a weighted sum of all the tree outputs:ω

$$b(x) = \sum_{j=1)}^{M} \omega_j c_j\,(x)$$

where $\omega_j \in R$ is the weight associated to tree j. In their paper Lou *et al.* (2017) provide agnostic tools, see also Freitas (2014), and Ribeiro *et al.* (2016a).

Local approach

The global approach is also relatively limited due to the complexity of the models to explain, thus recent research have focused on specific local explanation providing explainability for the prediction. A common method is based on visualization. The technique is used to explain CNN process when they recognize images. The method is based on salient mask which is a part of an image or a sentence in a text. The explanation c_l of the prediction is provided through a visualization of the area of an image for instance. A technique consists in assigning a relevance score for each layer backpropagating the effect of a decision on a certain image up to the input level. The function f used to extract the local explanation c_l is always not generalisable and strictly tied with the convolutional neural network.

Thus, **agnostic** approach has also been developed for the local approach. The agnostic solutions proposed for the outcome explanation problem implements function f such that **any type** of black box b can be explained. All these approaches are generalizable by definition and return a comprehensible local predictor c_l. The more popular technique uses again the concept of additive models to weight the importance of the features of the input dataset. It provides a graphical explanation of the decision process by visualizing the feature importance for the decisions, the capability to speculate on the effect of changes to the data, and the capability, wherever possible, to drill down and audit the source of the evidence, some examples are in Poulin *et al.* (2006).

A way to find f providing a local interpretable predictor c_l is known as the Local Interpretable Model-agnostic Explanations (LIME) approach. LIME approach for f does not depend on the type of data, nor on the type of black box b to be opened, nor on a particular type of comprehensible local predictor c_l, thus LIME is model-**agnostic** in its philosophy. The main intuition of LIME is that the explanation may be derived locally from the inputs generated randomly in the neighborhood of the input x to be explained, and weighted according to their proximity to it. Linear models as comprehensible local predictor c_l are considered returning the importance of the features, and as black box b, classifiers like decision trees, logistic regression, nearest neighbors, SVM, and random forest are usually tested. A weak point of this approach is the required transformation of any type of data in a binary format (Ribeiro, 2016a). Other local interpretability methods have been developed like the counter factual explanations, but with a different objective (Ribeiro, 2016b). We detail now these two main local approaches.

Local interpretable solutions

Coming back to the previous discussion it can exist a trade-off between the performance of the model and the effort required to interpret it - especially in complex domains like text and image analysis, where the input space is very large. In these contexts, accuracy is usually sacrificed for models that are enough compact and transparent to be comprehensible by humans. To try to associate accuracy and transparency local approaches can be considered and they will permit to answer to the objective assignated to this paper: '**why'** and '**how'** a decision is taken using ML algorithms.

LIME method

The Local Interpretable Model-agnostic Explanations (LIME) method interprets individual model predictions based on locally approximating the model around a given prediction. LIME refers to simplified inputs x' as "interpretable inputs," and introduce a mapping $x = hx\ (x')$ which converts a binary vector of interpretable inputs into the original input space. Different types of h_x mappings are used for different input spaces. For bag of words text features, h_x converts a vector of 1's or 0's (present or not) into the original word count if the simplified input is one, or zero if the simplified input is zero. For images, h_x treats the image as a set of super pixels; it then maps 1 to leaving the super pixel as its original value and 0 to replacing the super pixel with an average of neighboring pixels (this is meant to represent being missing).

The important point concerning this approach is that it is based on local accuracy meaning that when approximating the original model classifier b for a specific input x, local accuracy requires the explanation model c_l to at least match the output of b for the simplified input x' (which corresponds to the original input x).

If we denote l a measure of how unfaithful c_l is in approximating b in the locality around x', measured by a local kernel π_x, we need to minimize l(b, c_l, π_x) while having Ω(C) be low enough to be interpretable by humans. The explanation f(x') produced by LIME is obtained solving

$$f(x') = argmin_{cl \in C}\ l(b, c_l, \pi_x) + \Omega(C).$$

If we use an additive model as explanation model for b in the framework proposed by LIME where a mapping x = $h_x(x')$ converts a binary vector of interpretable inputs into the original input space, then the local approximation c_i tries to ensure that $c_i(z') \approx b(h_x(z'))$ whatever z' ≈ x' (note that $h_x(x') = x$ even though x' may contain less information than x because h_x is specific to the current input x).

An explanation solution that is linear function of binary variables is provided by:

$$c_i(z') = \varphi_0 + \sum_{(1=i)}^{M} \varphi_i z_i$$

where $z' \in [0,1]^M$, M is the number of simplified input variables, and $\varphi_0 \in R$. This explanation solution attributes an effect φ_i to each input, and summing the effects of all input attributions approximate the output b(x) of the original model. Some applications are provided in Ribeiro (2016a). This approach provides a local answer for the interpretation or the understanding of the model b ('**how**'): it tries to illustrate the way by which the predictions have been provided.

Unconditional counterfactual explanation method

This approach provides a way to understand how a given decision has been obtained, and can provide grounds to contest it, and advice on how the data input can change his or her behaviour or situation to possibly receive a desired decision (*e.g.* loan approval) in the future. This knowledge can be associated to the works implied in the fairness of machine learning development. Suppose that you were denied a loan because your annual income was 30,000 euros. If your income had been 45,000 euros, you would have been offered a loan? Here the statement of decision is followed by a counterfactual, or statement of how the world would have to be different for a desirable outcome to occur. The counterfactual approach aims to create the smallest possible change to obtain a desirable result. The difficulty is to have knowledge of the relevance of all the factors at play and of their possible change. Hence the idea of having a counterfactual explanation which modifies the values from which we start as little as possible. Thus, the underlying idea is quite simple: the idea is to find a neighboor from the input which provides a different prediction with the same classifier.

As before denote X the input, Y the output, and the classifier b_ω trained by finding the optimal set of weights ω_i that minimises an objective loss function l(.) over a set of input data X, then the objective is to compute:

$$argmin_\omega\ l(b_\omega(x), y) + \sigma(\omega)$$

where σ is a regularizer over the weights. The idea is to find a counterfactual x' as close to the original point x as possible such that b_ω (x') is equal to a new target y'. We can find x' by holding ω fixed and minimizing the related objective:

$$argmin_{x'}\ max_\lambda\ \lambda(b_\omega(x') - y')^2 + d(x,x')$$

where d(,) is a distance function that measures how far the counterfactual x' and the original data point x are from one another. In practice, maximisation over λ is done by iteratively solving for x' and increasing λ until a sufficiently close solution is found. The choice of the distance d is important, that of λ is less so. Depending on the data, the distance could be the L^1 or L^2 norm, the Manhattan distance weighted by the inverse median absolute deviation. As local minima are a concern, one can initialize each run with different random values for x' and select as counterfactual the best minimizer of the previous equation. These different minima can be used as a diverse set of multiple counterfactuals.

Thus, with this approach, the original classifier does not change, the inputs are concerned in the sense that for a given classifier they determine the prediction. Thus the method tries to answer to the question of **why** we got these predictions. In that sense counterfactuals represent an easy first step that balances transparency, explainability, and accountability with other interests such as minimising the regulatory burden on business interest or preserving the privacy of others,

while potentially increasing public acceptance of automatic decisions. Thus, it may prove a highly useful mechanism to meet the explicit requirements and background aims of the GDPR, but it is important to have in mind that this approach may be the target of specific attacks such as those known as 'adversarial attacks'.

How to compute explanations of model prediction?

In the previous paragraph we provided some strategies to interpret ML algorithms and their predictions without quantifying the approximations which have been proposed. Some works tend to answer to this problematic.

Global feature attribution is represented in the literature by several methods: gain, split count, and feature permutation. Gain is the total reduction of loss or impurity contributed by all splits for a given feature. Though its motivation is largely heuristic, gain is widely used as the basis for feature selection methods (Friedman *et al.,* 2011). Split Count consists in simply to count how many times a feature is used to split. Since feature splits are chosen to be the most informative, this can represent a feature's importance (Chen & Guestrin, 2016). With permutation one randomly permutes the values of a feature in the test set and then observes the change in the model's error. If a feature's value is important then permuting it should create a large increase in the model's error (Auret & Aldrich, 2011). Two other measures have been also developed: the SHAP measures and the Quantitative Input Influence measures (QII).

SHAP values for LIME approach

The ability to correctly interpret a prediction model's output is extremely important but the solutions are not simple and it exist several strategies. These strategies engender appropriate user trust, provide insight into how a model may be improved, and support understanding of the process being modeled. In some applications, simple models (*e.g.*, linear models) are often preferred for their ease of interpretation, even if they may be less accurate than complex ones. It could be interesting to compare the accuracy of the 'interpretable' built when blackbox models are used. It is the objective of the SHapley Additive exPlanation (SHAP) values which are based on a unification of ideas from game theory (Strumbelj & Kononenko, 2014), and local explanations (Ribeiro, 2016a).

We have seen previously that the more understanding solutions are based on families of additive models. In the case of the local approximation proposed by the LIME method, the solution is provided by equation introduced in subsection 5-1. The explanation model $c_l\ (x')$ matches the original model b(x) when x = $h_x(x')$, and φ_0 = b(h_x(0)) represents the model output with all simplified inputs toggled off (*e.g.* missing). As soon as an additive interpretable model is built, the SHAP values characterize the additive inputs of this interpretable model. For instance if the additive model reduces to the very simple model:

$$c_l(x) = b(x) = \sum_{(j=1)}^{M} \omega_j x_j + a$$

The SHAP values are equal to φ_0 (x) = a and φ_j(x) = ω_j (x_j- E[x_j]). The exact computation of SHAP values is challenging, some examples can be found in Lundberg and Lee (2017) and Strumbelj and Kononenko (2014).

In case of a more general additive interpretable model like the model introduced in subsection 5-1, it has been shown that this model presents three interesting properties: local accuracy, missingness, and consistency: the local accuracy states that the sum of the feature attributions is equal to the output of the function we are seeking to explain; missingness states that features that are already missing (such that z= 0) are attributed no importance; consistency states that changing a model so a feature has a larger impact on the model will never decrease the attribution assigned to that

feature. The evaluation of the effect of missing features has on a model c_I is done through the use of the function h_x evaluating b(h_x(z)) and calculate the effect of observing or not observing a feature (by setting z= 1 or z = 0).

To compute the SHAP values for the approximation done through the linear model, we define $b_x(S)$ = b(h_x(z')) = E[b(x) | x_S] where S is the set of non zero indexes in z', and E[b(x) |x_S] is the expected value of the function conditioned on the inputs variables x_S belonging to S. SHAP values attribute φ_j values to each variable:

$$\varphi_j = \sum_{(SCN-\{i\})} \frac{|S|!(M - |S| - 1)!/M!}{M!} \; [b_x(S \cup \{i\}) - b_x(S)]$$

Where N is the set of all inputs.

The computational complexity of SHAP values have been extensively studied and some ways to reduce its computational times are discussed in Lundberg *et al.* (2018), Chen *et al.* (2019), and Erion *et al.* (2019).

QII and Counterfactual approach

The Quantitative Input Influence measures (QII) model the difference in the quantity of interest when the system operates over two related input distributions: the real distribution and a hypothetical (or counterfactual) distribution that is constructed from the real distribution in a specific way to account for correlations among inputs. Specifically, if we are interested in measuring the influence of an input on a quantity of interest of the system behavior, we construct the hypothetical distribution by retaining the marginal distribution over all other inputs and sampling the input of interest from its prior distribution. This choice breaks the correlations between this input and all other inputs and thus lets measure the influence of this input on the quantity of interest, independently of other correlated inputs.

Using the same notations as before. As QII quantifies the use of an input for individual outcomes, this quantity is defined for a particular individual. Denote x this individual and c the classifier retains in fine, the quantity E[c(.) = 1 | X=x] represents the expectation of the classifier c evaluating to 1 for the individual x. The influence measure, when the positive classification is the objective is computed as

$$\text{QII}(x_i) = \text{E}[c(X) = 1 \mid X=x] - \text{E}[c(XU_{xi})) = 1 \mid X=x]$$

where the random variable XU_{xi} corresponds to a randomized intervention on input x which is replaced with a random sample x_i. Thus, we have switched between the original distribution, represented by the random variable X, and the intervened distribution represented by XU_{xi}, Datta *et al.* (2016) for some applications.

As an example consider an analyst who asks: "What is the influence of the input gender on positive classification for women?" If it observes that 20% of women are approved according to his classifier, then, he replaces every woman's field for gender with a random value, and if he notices that the number of women approved does not change, this means that an intervention on the gender variable seems not causing a significant change in the classification outcome. Now the analyst can repeat the same process with 'weight lifting ability' variable and if the results show a 20% increase in women's hiring, therefore he can conclude that for this classifier, the variable 'weight lifting ability' has more influence on positive classification for women than gender. By breaking correlations between gender and weight lifting ability, can be a way to establish a causal relationship between the outcome of the classifier and the inputs. These facts are interesting and need to be more developed to verify what the correlations are really suppressed, and to understand the causal relationship between classifier's output and inputs.

Conclusion

The governance of decision making is an important task with developement of AI in industies and banking system. The requirement for explanation, a requirement codified through risk management in traditional sectors of industry and by the rules of certain professions (medicine, law), is also present in the AI sphere, where certain aspects are covered by legislation (for instance RGPD law).

To explain an algorithm enabling its users to understand what it does, with enough details and arguments to instill trust is a difficult task. The global and local solutions discussed here provided very interesting pistes, in particular the local counter-factual approach which could be a method from which the regulator could draw inspiration to verify the accountability of the algorithms. But many risks exist: one of them related to the interpretability of the models is known under the name of adversarial attacks and its study is in full expansion, we do not discuss it in this paper being beyond the objective of it, nevertheless some recent and interesting references are Dylan *et al.* (2018), Kim and Malde (2020), and Slack *et al.* (2020), for a review Bogroff and Guégan (2019).

In summary, for an algorithm to be explainable, its principles must be sufficiently documented to be comprehensible to all users; the transition from algorithm to code, then the execution of the program, must be formally verified. Ultimately, the explainability of an algorithm relies on rigorous methods, but also on a body of unformalized knowledge shared between human beings. As a result, a compromise has to be found between learning capacities and explainability. This compromise needs to be evaluated in relation to the field of application: while explainability is not in principle essential in applications such as games, it is crucial once the interests, rights or safety of people are concerned. From a social point of view people do not ask why an event P happened, but rather why event P happened instead of some event Q (Papernot, 2018; Alvarez-Melis, 2018).

It exist several works coming from the legal literature which propose procedures for the transparency of the source code in case of auditing purpose but also for the understanding of the users, making software verification, fairness random choices, disclosing commitments. They suggest the developers to publish in advance commitments explaining how the systems do without disclosing how those systems work up front. These procedures are complementary to the previous analysis and preaches for the need of platforms and algorithms to be evaluated (compliance, fairness, trustworthiness, neutrality, transparency…). This will contribute to good algorithm governance. In a complete paper, Kroll *et al.* (2016) provide solutions to make accountable machine learning algorithms, recommandations are also done in Cerna (2018), Linkov *et al.* (2018). Rules, regulation and governance are addressed in Barocas *et al.* (2013).

Ultimately, the role of humans must be preserved throughout the process of explicability, which requires significant expertise from those in charge of systems using AI in decision-making processes.

References

AGARWAL A., BEYGELZIMER A., DUDIK M., LANGFORD J. & WALLACH H. (2018), "A reductions approach to fair classification", arXiv preprint arXiv:1803.02453.

BIRAN O. & COTTON C. (2017), “Explanation and justification in machine learning: A survey”, in IJCAI-17 workshop on explainable AI (XAI), Vol. 8, No. 1, pp. 8-13.

BAROCAS S., HOOD S. & ZIEWITZ M. (2013), “Governing algorithms: A provocation piece”, Available at SSRN 2245322.

BERK R., HEIDARI H., JABBARI S., KEARNS M. & ROTH A. (2018), "Fairness in criminal justice risk assessments: The state of the art. Sociological Methods and Research", online publication, https://doi.org/10.1177/0049124118782533 %0049124118782533.

BOGROFF A. & GUEGAN D. (2019), "Artificial Intelligence, Data, Ethics An Holistic Approach for Risks and Regulation", University Ca'Foscari of Venice, Dept. of Economics Research Paper Series, (19).

BURRELL J. (2016), "How the machine 'thinks': Understanding opacity in machine learning algorithms", *Big Data and Society,* 3(1), 2053951715622512.

Cerna Collectif (2018), "Research Ethics in Machine Learning" [Research Report] CERNA; ALLISTENE, pp.51, hal-01724307.

CHEN H., LUNDBERG S. & LEE S. I. (2019), "Explaining Models by Propagating Shapley Values of Local Components", arXiv preprint arXiv:1911.11888.

CHZHEN E., DENIS C., HEBIRI M., ONETO L. & PONTIL M. (2020), "Fair Regression via Plug-in Estimator and Recalibration With Statistical Guarantees", hal-02501190, https://hal.archives-ouvertes.fr/hal-02501190.

CRAVEN M. W. (1996), "Extracting comprehensible models from trained neural networks", University of Wisconsin-Madison Department of Computer Sciences.

DATTA A., SEN S. & ZICK Y. (2016), "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems", in 2016 IEEE symposium on security and privacy (SP), pp. 598-617, IEEE.

DOSHI-VELEZ F. & KIM B. (2017), "Towards a rigorous science of interpretable machine learning", arXiv preprint arXiv:1702.08608.

DWORK C., HARDT Y., PITASSI T., REINGOLD O. & ZEMEL R. (2012), "Fairness Through Awareness", *Proceedings* of the 3rd Innovations of Theoretical Computer Science, 214 -226.

DYLAN H., GIOE D. V. & GOODMAN M. S. (2018), "The CIA and the pursuit of Security", Middle East, 1948, 63.

ERION G., JANIZEK J. D., STURMFELS P., LUNDBERG S., LEE S.-I. (2019), "Learning explainable models using attribution priors", arXiv preprint arXiv:1906.10670.

FELDMAN M., FRIEDLER S. A., MOELLER J., SCHEIDEGGER C. & VENKATASUBRAMANIAN S. (2015), "Certifying and removing disparate impact", *Proceedings* of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 259-268.

FERMANIAN J.-D. & GUEGAN D. (2020), "Fair learning with bagging", Working paper.

FREITAS A. A. (2014), "Comprehensible classification models: A position paper", SIGKDD Explor. Newsl., 15(1), 1–10, ISSN 1931-0145.

GOODMAN B. & FLAXMAN S. (2017), "European Union regulations on algorithmic decision-making and a 'right to explanation'", *AI magazine,* 38(3), 50-57.

GUIDOTTI R., MONREALE A., RUGGIERI S., TURINI F., GIANNOTTI F. & PEDRESCHI D. (2018), "A survey of methods for explaining black box models", ACM computing surveys (CSUR), 51(5), 1-42.

HARDT M., PRICE E. & SREBRO N. (2016), "Equality of opportunity in supervised learning", Advances in neural information processing systems, 3315-3323.

KAMIRAN F. & CALDERS T. (2012), "Data preprocessing techniques for classification without discrimination", *Knowledge and Information Systems,* 33(1): 1-33.

KIM H. & MALDE K. (2020), "Proper measure for adversarial robustness", arXiv preprint arXiv:2005.02540.

KROLL J. A., BAROCAS S., FELTEN E. W., REIDENBERG J. R., ROBINSON D. G. & YU H. (2016), "Accountable algorithms", U. Pa. L. Rev., 165, 633.

KUSNER M. J., LOFTUS J., RUSSELL C. & SILVA R. (2017), "Counterfactual fairness", Advances in Neural Information Processing Systems, (NeurIPS 2017), 4066-4076.

LINKOV I., TRUMP B. D., POINSATTE-JONES K. & FLORIN M. V. (2018), "Governance strategies for a sustainable digital world", *Sustainability,* 10(2), 440.

LIPTON Z. C. (2018), "The mythos of model interpretability", *Queue,* 16(3), 31-57.

LOU Y., CARUANA R. & GEHRKE J. (2012), "Intelligible models for classification and regression", in *Proceedings* of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 150-158).

LU Q., CUI Z., CHEN Y. & CHEN X. (2017), "Extracting optimal actionable plans from additive tree models", *Frontiers of Compututational Science,* 11(1): 1-15.

LUNDBERG S. M. & LEE S. I. (2017), "A unified approach to interpreting model predictions", in Advances in neural information processing systems, pp. 4765-4774.

LUNDBERG S. M., ERION G. G. & LEE S. I. (2018), "Consistent individualized feature attribution for tree ensembles". arXiv preprint arXiv:1802.03888.

MELIS D. A. & JAAKKOLA T. (2018), "Towards robust interpretability with self-explaining neural networks", in Advances in Neural Information Processing Systems, pp. 7775-7784.

MILLER T. (2019), "Explanation in artificial intelligence: Insights from the social sciences", *Artificial Intelligence,* 267, 1-38.

Organization for Economic Co-Operation and Development (OECD), (2016), "OECD Science, Technology, and Innovation Outlook 2016", OECD Publishing: Paris, France, Chapter 2, pp. 110–111.

PAPERNOT N. (2018), "A Marauder's Map of Security and Privacy in Machine Learning", arXiv preprint arXiv:1811.01134.

POULIN B., EISNER R., SZAFRON D., LU P., GREINER R., WISHART D. S. & ANVIK J. *et al.* (2006), "Visual explanation of evidence with additive classifiers", in *Proceedings* of the National Conference on Artificial Intelligence, Vol. 21, No. 2, p. 1822, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press.

RIBEIRO M. T., SINGH S. & GUESTRIN C. (2016a), "Why should I trust you? Explaining the predictions of any classifier", in *Proceedings* of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135-1144.

RIBEIRO M. T., SINGH S. & GUESTRIN C. (2016b), "Model-agnostic interpretability of machine learning", arXiv preprint arXiv:1606.05386.

SLACK D., HILGARD S., JIA E., SINGH S. & LAKKARAJU H. (2020), "Fooling lime and shap: Adversarial attacks on post hoc explanation methods", in *Proceedings* of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 180-186.

ŠTRUMBELJ E. & KONONENKO I. (2014), "Explaining prediction models and individual predictions with feature contributions", *Knowledge and information systems,* 41(3), 647-665.

WADSWORTH C., VERA F. & PIECH C. (2018), "Achieving fairness through adversarial learning: an application to recidivism prediction", arXiv preprint arXiv:1807.00199.

WACHTER S., MITTELSTADT B. & RUSSELL C. (2017), "Counterfactual explanations without opening the black box: Automated decisions and the GDPR", *Harv. JL and Tech.,* 31, 841.

ZAFAR M. B., VALERA I., GOMEZ RODRIGUEZ M. & GUMMADI K. P. (2017), "Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment", in *Proceedings* of the 26th international conference on world wide web, pp. 1171-1180, April.

ZEMEL R., WU Y., SWERSKY K., PITASSI T. & DWORK C. (2013), "Learning fair representations", International Conference on Machine Learning, 325-333.

ZLIOBAITE I. (2015), "On the relation between accuracy and fairness in binary classification", arXiv preprint arXiv:1505.05723.

Intelligence artificielle et contrôle de gestion : un rapport aux chiffres revisité et des enjeux organisationnels

Par **Christian MOINARD**
Professeur à Audencia
et **Nicolas BERLAND**
Professeur à l'Université Paris-Dauphine

Introduction

Le développement du *big data* et des données issues des réseaux sociaux donnent aux contrôleurs de nouveaux terrains de jeux pour comprendre la performance des entreprises. La possibilité d'exploiter ces données *via* des algorithmes et des programmes d'intelligence artificielle ouvre un nouveau paradigme pour les technologies et pratiques du pilotage et du contrôle de gestion. L'accès à cette « informatique cognitive » est de nature à changer les activités des contrôleurs (Sponem, 2018), la répartition du travail contrôleurs/managers et les organisations elles-mêmes.

Les deux développements technologiques décrits ci-dessus, intimement liés, ont des conséquences sur les individus et sur les organisations. Dans le premier cas, c'est le rapport que les agents entretiennent avec la performance chiffrée qui se trouve modifié, et, partant, le métier de contrôleur. Mais ce changement individuel s'accompagne d'un changement organisationnel, si on veut profiter à plein des effets de ce nouveau paradigme du contrôle.

Un rapport aux données chiffrées revisité

Avec le développement du *big data* et de l'intelligence artificielle (IA), le rapport aux chiffres des décideurs est modifié. Par nature, les chiffres, et spécialement ceux utilisés dans le management, sont des conventions (Boussard, 1998). Ils intègrent des hypothèses et des raccourcis quant à la manière dont ils sont collectés, agrégés et interprétés. Trop peu d'acteurs sont conscients de ces conventions, de ces constructions sociales, et beaucoup prennent les chiffres pour argent comptant. Ceux familiers de ces conventions ont parfois tendance à rejeter les chiffres, les estimant truqués, ou perdent confiance dans leur valeur. Les théories du complot ne sont jamais loin. Bien peu vont prendre les arguments numériques avec prudence, voire sagesse, sans les rejeter pour autant. *Big data* et IA risquent de ne faire que renforcer ces tendances.

Une collecte des chiffres modifiée

Le *big data* et l'IA ont le potentiel de transformer en boite noire ce qui hier (et sans doute encore pour un certain temps) tenait sur une feuille Excel ou était issu d'un logiciel ERP (*Entreprise Ressource Planning),* dont la nature transactionnelle permettait un traçage des données. On peut « voir » les données, remonter à leur source, bref les mettre en question. De même, leur traitement est l'affaire d'individus qui utilisent des heuristiques d'harmonisation, d'agrégation et d'interprétation, souvent critiquables, mais pour lesquels il est possible de demander des comptes.

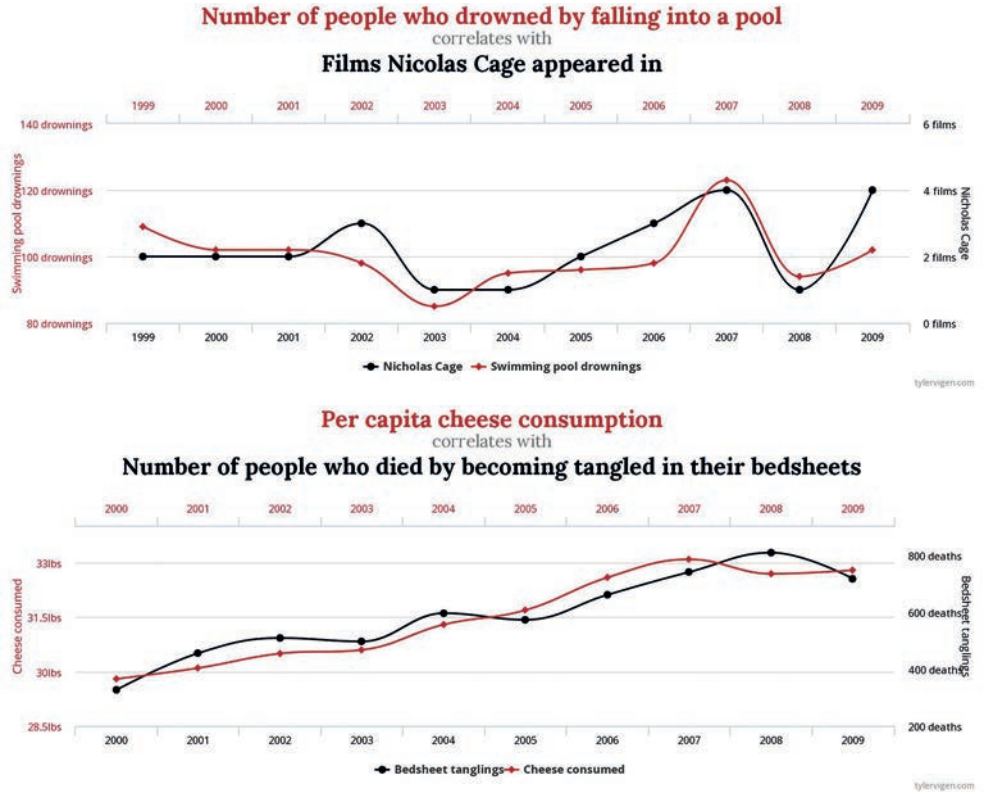
Par nature, les données du *big data*, sont plus compliquées à appréhender. Elles changent rapidement, sont trop nombreuses pour être visualisées dans le "*data lake"*, dépendent de la perspective depuis

laquelle on les regarde (depuis quel "*lake shore*"), sont souvent en doublons (un collègue du Lamsade rappelait récemment dans un *workshop* interne à Dauphine que de nombreux congrès scientifiques ont pour objet le traitement des doublons, signe que cela est plus complexe que sur Excel). Ces données sont moins structurées, mais sans doute aussi plus intéressantes, car elles portent en elles l'ambiguïté des situations. Mais ces données deviennent des boites noires, riches et incertaines.

De plus, avec l'IA, un algorithme remplace l'interprétation humaine. Outre que cet algorithme porte les croyances de ses concepteurs, il semble qu'il puisse se développer de manière autonome. Un humain peut être amené à se justifier, mais pas un algorithme. Aussi, qui porte alors la responsabilité des décisions ? Comment peut-on faire confiance à une décision dont on ne comprend pas les bases ? Comment remplacer la confiance interpersonnelle, base des croyances dans les interprétations de l'autre ?

Une interprétation des données sous tension

L'interprétation est elle aussi modifiée. L'algorithme produit essentiellement des corrélations là où le décideur a besoin de causalités. Certes, bien souvent, le décideur agit sur la base de corrélations incertaines (la mise en correspondance de deux séries de chiffres) et donne du sens à cette relation. Mais parce que c'est un être humain, le lecteur des chiffres sera souvent plus suspicieux sur les croyances qui sous-tendent l'interprétation. Un algorithme risque fort d'être plus souvent cru dans une espèce de fantasme technologique, surtout si dans les premières interactions il a été sans faille. Mais dès que la répétitivité de l'usage sera lancée, il deviendra vite une boite noire. Par ailleurs, l'algorithme produisant des corrélations entre des variables incertaines, les décideurs risquent de passer beaucoup de temps à interpréter le type de graphiques ci-dessous avant de s'apercevoir de l'incohérence de telles corrélations (Calude et Longo, 2017).



Source : https://www.tylervigen.com/spurious-correlations

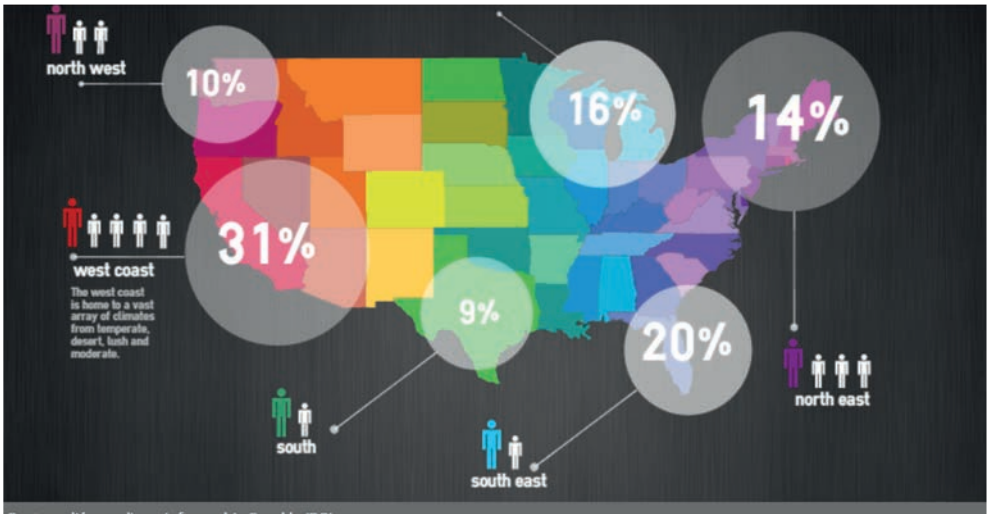
Le temps passé à l'analyse de corrélations risque d'être d'autant plus important que la nécessité d'alimenter des algorithmes pousse à une multiplication des sources de données et des systèmes de traitement. Les décideurs doivent ainsi composer avec des *"packages"* de systèmes d'information, c'est-à-dire des ensembles dont la cohérence interne n'est pas totalement assurée. Cette multiplication des données génère un temps passé à comprendre les interactions entre données et à analyser les contradictions, et complexifie la prise de décision. Le décideur se trouve pris dans les méandres d'une double contradiction :

- renforcer le volume et la nature des données pour assurer la qualité des systèmes de prédiction,
- simplifier et réduire les données, et rendre l'exercice de décision acceptable avec la temporalité de l'action.

Il faut alors former les décideurs à comprendre la différence entre une corrélation et une causalité (pas si simple compte tenu de l'incertitude philosophique entourant la notion de causalité), mais aussi à optimiser les flux de données utiles à la prise de décision. Pour cela, il est nécessaire « de donner du sens » aux chiffres. À chaque situation de gestion appartient son propre processus de décision. Cet exercice de recherche de sens demande de construire des histoires communes en lien avec les activités, mais aussi à apprendre à en débattre. En d'autres mots, le besoin de renforcer l'interprétation des chiffres est fort et doit être un point dominant de la formation (encore plus qu'avant).

Le périmètre du métier de contrôleur remis en question

La bonne interprétation des résumés de la réalité que sont les chiffres nécessite une connaissance forte des modèles d'affaires à gérer. Si le manager a, on peut l'espérer, une bonne connaissance de son métier, ce n'est pas toujours aussi évident pour le contrôleur (l'homme/la femme des chiffres) qui préfère parfois davantage Excel à une connaissance étroite du métier. Certains y arrivent très bien (et finissent souvent par quitter la fonction). Mais on peut se demander si le contrôleur ne doit pas être un poste de transition pour des managers en pleine progression de carrière. Débarrassé des aspects fastidieux de la collecte et de la mise en ordre des données, il deviendrait l'expert de l'interprétation, au côté d'un *data analyst* qui produit les chiffres en interface avec les algorithmes. À l'inverse, on peut aussi se demander si le contrôleur ne doit pas devenir un *data savvy* ou *data sterwardship*, un sachant de la donnée, capable de comprendre d'un point de vue des opérations ce qu'il y a dans ces boites noires. S'il n'a pas forcément la compétence technique, il sait, au moins, crédibiliser les chiffres.



Capture d'écran d'une infographie Easel.ly (DR)

Enfin, ces nouvelles opportunités renforcent le besoin de rendre attrayants les chiffres. À la fois parce que nous sommes entrés dans une ère du *design* (ce qu'on produit doit être beau) et parce que le volume de données va croissant, il faut qu'ils soient faciles à lire, à comprendre, plaisants à regarder. Des outils comme Power BI travaillent sur la *datavisualisation*, car elle est aussi une condition de la bonne utilisation des données chiffrées.

Mais cela demande aussi de faire cohabiter différentes représentations de la donnée, c'est-à-dire différents cadres permettant le passage d'une vision micro (l'analyse des événements) à une vision macro (l'identification des tendances lourdes). Si le contrôleur de gestion est passé maitre dans sa capacité à faire cohabiter différents référentiels de données (comptables, outils de gestion, opérationnel…), son rôle va se trouver bousculé par la montée en puissance des modèles issus de l'informatique cognitive, auxquels il n'est pas accoutumé.

Une organisation adaptée

Les conséquences de l'introduction du *big data* et de l'IA ne se limitent sans doute pas aux seuls individus. Si de nouvelles pratiques sont possibles, pour donner leur plein potentiel, les organisations doivent être modifiées. Cela concerne tout à la fois les structures *corporate* mais aussi les modalités d'ajustement des individus (voir sur le sujet le dossier de la DFCG sur la transformation de la fonction financière).

De nouvelles pratiques… en devenir

Le *Big data* et l'IA devraient permettre d'élargir le mix de contrôle ou du moins de le transformer. Si les étapes de description et de diagnostic sont bien maitrisées par les professionnels, les nouveaux systèmes de l'informatique cognitive devraient permettre de les systématiser. La phase descriptive a été beaucoup améliorée par la *business intelligence*, bien qu'il soit possible d'aller plus loin (information obtenue en mode conversationnel par exemple). La partie diagnostique, au moins pour les tâches courantes, a été améliorée (même si des progrès restent à faire) en systématisant les alertes (la détection de fraudes par exemple).

Mais c'est sur la partie « prédiction et prescription » que les choses devraient le plus changer. La prévision reste encore le parent pauvre des tâches de contrôle. Or, si « gérer, c'est prévoir », les managers pourraient gagner en efficacité, car ces systèmes d'information seront sans doute utiles pour améliorer les modélisations du futur de l'entreprise. En repérant des signaux faibles,

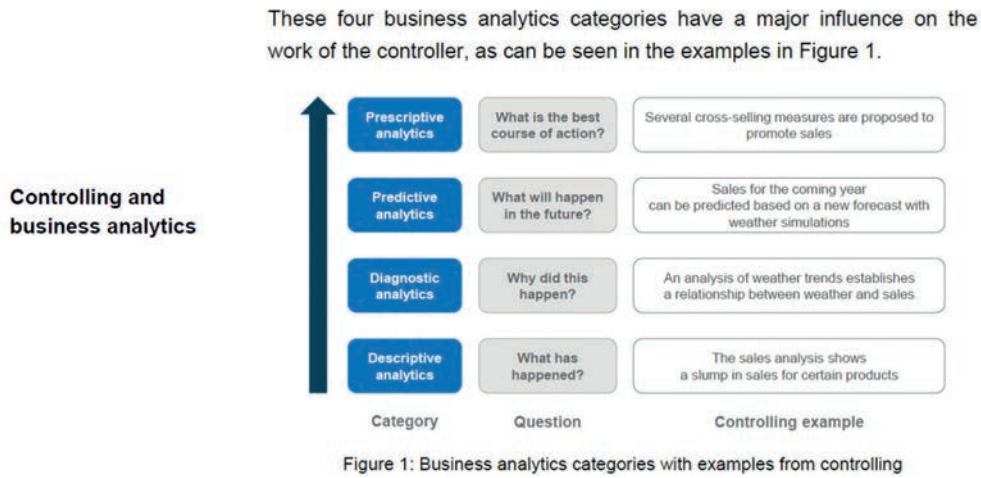


Figure 1: Business analytics categories with examples from controlling

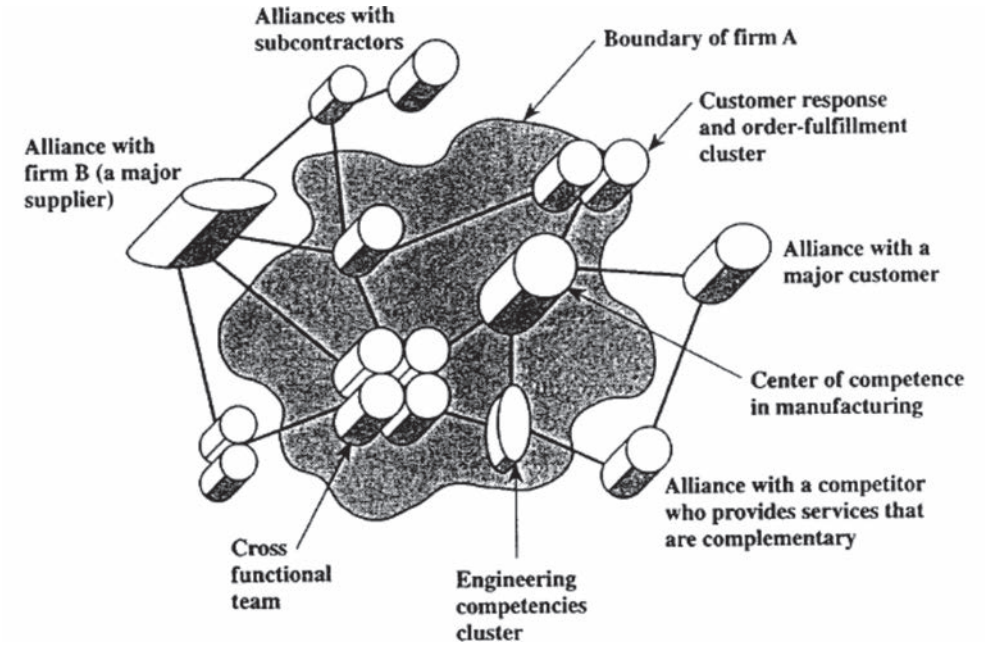
des infléchissements, en tenant compte de la position des concurrents, les entreprises pourraient gagner en visibilité, sans le *monitoring* constant d'un être humain.

Mais c'est dans le domaine de la prescription que les attendus peuvent tout changer. Par exemple, que ce soit en *BtoC*, avec la capacité des algorithmes à proposer de nouveaux produits aux consommateurs et donc ainsi d'orienter ses choix, ou en *BtoB,* avec la maintenance préventive qui suggère des besoins de réparation, le développement du *big data* et de l'IA peut tout à la fois participer au développement des marchés et à la réduction des variances sur les prévisions. Tout reste à faire, toutefois, encore aujourd'hui.

Le changement des structures *corporate* : le passage à la forme en N

L'informatique cognitive devrait aussi changer les modes d'organisation et de structuration des entreprises. L'histoire des entreprises montre que les avancées en termes d'outils ou de pratiques, notamment en contrôle, se sont accompagnées d'un changement dans les formes structurelles (Berland, 2020). L'introduction du calcul des coûts s'est ainsi accompagnée de l'adoption de la forme en U (forme fonctionnelle), le budget et les tableaux de bord se sont accompagnés de l'adoption de la forme en M (pour multidivisionnelle). L'insistance sur pilotage financier a généralisé les *holdings* (*H-form)*.

L'introduction de l'informatique cognitive semble aller de pair avec le développement de forme en N (pour *network* ou réseaux), les *N-form/cluster organizations.* L'informatique cognitive n'est pas seule responsable de ce changement qui semble plus systémique. Le passage d'une économie de production et de services à une économie de la connaissance semble avoir conduit nombre d'entreprises à modifier leur organisation pour adopter des modes plus flexibles ou agiles.



N-form/cluster organizations, self-designing organizations, information-based organizations, and post-industrial organizations

L'informatique cognitive ne fait que renforcer cette transition. L'information a moins besoin d'être structurée que dans une organisation « normale ». La division du travail s'y fait sur la base des

connaissances, et donc de l'information qui est au cœur de l'informatique cognitive. Cela permet une décentralisation accrue des organisations (baisse des coûts de transaction pour Williamson). Et le mode de coordination à l'intérieur de ces structures se fait par projet, comme nous allons le voir ci-dessous.

Des modes de coordination moins « silotées »

À l'intérieur des organisations (si on peut encore parler d'intérieur et d'extérieur…), les modes de coordination semblent aussi touchés. L'informatique transactionnelle s'est développée en même temps que le *réengineering* et la gestion par processus. Si la couche informatique que constituaient les ERP n'était pas suffisante à elle seule pour améliorer le fonctionnement des organisations (Meysonnier et Pourtier, 2005) [(1)], il semble que le développement de l'informatique cognitive s'accompagne d'autres modes de coordination (qui là aussi ne dépendent pas que d'elle), comme le *lean management*, l'holacratie, les adhocraties ou encore les entreprises libérées. Si les modèles diffèrent, le message est toujours le même. Il s'agit de tenter de « dessiloter » des organisations devenues trop bureaucratiques. Mais cela change la manière de piloter les organisations. Ce mouvement vers une plus grande décentralisation limite la vision de l'exhaustivité des actions par un seul acteur. La mise en cohérence des différentes actions et la stabilisation des processus n'est plus faite par la hiérarchie, mais par le transfert d'informations.

L'informatique cognitive participe au renforcement de cette agilité. En rendant accessibles de nouvelles informations, en programmant des étapes de validation et en partageant des relations de causalités (vérifiées ou supposées), elle participe à l'émergence de nouveaux schémas de coordination.

Cette évolution des structures organisationnelles, dans laquelle s'inscrit (plus qu'elle ne la provoque) l'informatique cognitive, influence les systèmes de contrôle et le rôle des contrôleurs de gestion. Les processus et organisations rendus moins modélisables, il devient aussi plus difficile d'appréhender la chaîne de causes et d'effets mettant en relation l'utilisation des ressources et les résultats attendus. Cette fonction de modélisation organisationnelle, à laquelle participent les contrôleurs de gestion, devrait être plus complexe. Face à de nouvelles modalités de représentation des organisations dont ils ne seront pas maîtres, les contrôleurs de gestion devront trouver une nouvelle légitimité.

Conclusion

Plus qu'une transformation technologique, le *big data*, l'intelligence artificielle et l'ensemble des technologies concourant au développement d'une « informatique cognitive » modifient les relations que nous entretenons avec les données, mais aussi les représentations que nous nous faisons des organisations. Ces transformations induisent et accompagnent de nouvelles modalités de coordinations organisationnelles. De nouveaux rôles pourraient apparaitre pour accompagner la mise en œuvre de ces nouvelles coordinations, à l'interface entre la maitrise des modèles mathématiques, l'analyse économique, la connaissance des activités et la compréhension des organisations. Quels formats ces fonctions prendront-elles, quels profils de compétences leur seront rattachés et quel sera la place des contrôleurs de gestion restent des questions actuellement sans réponse, mais cruciales pour cette profession.

(1) MEYSSONNIER F. & POURTIER F. (2005), « Les ERP changent-ils le contrôle de gestion ? Comptabilité et Connaissances », May, France.

Bibliographie

BERLAND N. (2020), *Le contrôle de gestion*, Que-sais-Je, Presses Universitaires de France, 128 p.

BOUSSARD V. (1998), « Les indicateurs de gestion comme construction sociale : l'exemple des CAF », *Revue des politiques sociales et familiales,* 54, pp. 51-61.

CALUDE C. & LONGO G. (2017), «The deluge of spurious correlation in big data, Foundations of science», 22, 3, pp. 595-612.

DFCG 2018, « IA, Big data, robotisation : quels impacts pour la direction financière ? », Cahiers techniques.

MEYSSONNIER F. & POURTIER F. (2005), « Les ERP changent-ils le contrôle de gestion ? Comptabilité et Connaissances », Congrès de l'Association Francophone de Comptabilité, May, France.

Gupta S., Karb A. K., Baabdullah A. & Al-Khowaiter W. A.A. (2018), "Big data with cognitive computing: A review for the future", *International Journal of Information Management,* Volume 42, p 78-89.

SPONEM S. (2018), « Une «société du contrôle» sans contrôle de gestion ? Réflexion sur le Big Data. », *in* Le libellio d'Aegis, Volume 14, n° 1, *Faire sens de l'évaluation – le cas du contrôle de gestion,* printemps.

Quelle régulation juridique pour l'intelligence artificielle ?

Par **Alain BENSOUSSAN**
Avocat au Barreau de Paris, Lexing Alain Bensoussan Avocats

L'intelligence artificielle (IA) va changer en profondeur notre façon de travailler et de voir le monde. De par leur utilisation généralisée, les technologies de l'intelligence artificielle (documents, données, algorithmes, robots…) vont devenir des technologies « ordinaires » en lieu et place des technologies existantes. Elles le sont d'ailleurs déjà très largement.

Dans le sillage de la transition numérique, la transition intelligente conduit les acteurs économiques à repenser leur modèle, à proposer des modèles disruptifs, à faire émerger de nouveaux services et évoluer leurs processus et leur organisation de façon à en améliorer la performance.

Seules les structures qui sauront s'approprier les solutions les plus créatives et les plus disruptives pourront tirer parti de cette révolution technologique.

À condition que soit encadrée sur les plans juridique et éthique cette révolution technologique, qui offre par ailleurs des perspectives sans précédent en termes de créativité et de *business.*

En tout état de cause, ces nouveaux modèles seront pilotés par le juridique dans de très nombreuses situations.

S'ajoute à ces enjeux de *business development* le fait que les attentes sociales sont, en la matière, immenses.

Faut-il avoir peur de l'IA ?

Lorsqu'un processus de décision, initialement confié à un agent humain, se trouve en tout ou partie automatisé, il est une évidence : les droits fondamentaux des personnes conservent toute leur intensité. La dignité doit être préservée, tandis que l'atteinte à la vie privée – si elle ne peut être évitée – doit être respectueuse du principe de proportionnalité, et l'absence de discrimination doit être assurée (sur ces questions, voir J. Bensoussan et J.-F. Henrotte, *Legal Aspects of Artificial Intelligence,* LexisNexis UIA, 2019).

En outre, les techniques d'intelligence artificielle sont « datavores » : leur performance dépend de la volumétrie et de la qualité de cette donnée, qui est souvent personnelle, et les fonctions sociales touchées peuvent être d'une grande sensibilité : la médecine, la justice, la sécurité ou l'emploi.

Cela met en tension les droits fondamentaux, pour lesquels un pilotage par les risques et une démarche centrée sur le rapport coût/avantage sont toujours plus délicats à implémenter.

Une démarche de régulation plus collaborative est donc utile, à l'instar des initiatives en matière « d'éthique de l'intelligence artificielle », dont l'objectif consiste à faire émerger des consensus conformes à nos système de valeurs et dont l'expression se retrouve dans les instruments juridiques à la normativité la plus élevée (constitution étatique, charte des droits fondamentaux de l'UE ou convention EDH… voir A. Bensoussan et J. Bensoussan, *Robot, IA et Droit,* Bruylant, 2019).

Le « droit souple », un instrument de régulation adéquat

Dans l'attente d'un encadrement juridique en devenir, cet avènement est le fruit de normes issues pour partie du « droit souple » ou *soft law*, parfois qualifié de « droit mou », qui regroupe un ensemble de règles de droit, tantôt obligatoires, tantôt non contraignantes, mais qui, en pratique, ont le mérite de « montrer le chemin » à une industrie du numérique dans l'attente de solutions, et à des utilisateurs d'objets connectés qui souhaitent être rassurés sur l'utilisation qui est faite de leurs données personnelles.

Résolutions, codes de bonne conduite, directives, *guidelines,* livres blancs, commissions et groupes de travail, sous l'égide d'organismes divers… C'est sur les bases de ce droit d'un genre nouveau que repose l'encadrement juridique du développement exponentiel des activités robotiques, des algorithmes et de l'IA.

Un « droit souple » a complété les premières décisions de jurisprudence rendues en la matière ainsi que les dispositions éparses disséminées dans des lois récentes à vocation générale.

Qu'on en juge :

- la Commission européenne a publié en juin 2019 une version mise à jour de ses lignes directrices visant à promouvoir une IA digne de confiance ;
- le Parlement européen a adopté en février 2019 une résolution sur une politique industrielle européenne globale sur l'intelligence artificielle et la robotique ;
- le Règlement européen sur la protection des données (RGPD), entré en application le 25 mai 2018, pose comme principe qu'une personne ne peut faire l'objet d'une décision fondée exclusivement sur un traitement automatisé ;
- le Conseil constitutionnel a encadré en juin 2018 le recours à des algorithmes autoapprenants (*deep learning*) comme fondement d'une décision administrative individuelle.

Autant de recommandations qui, si elles ne revêtent pas de nature obligatoire ou contraignante, posent les jalons d'un encadrement juridique de l'IA.

En 2019, face à de possibles incertitudes quant aux règles applicables, en matière de responsabilité civile et pénale, aux véhicules autonomes, le Gouvernement français a décidé de clarifier le régime de responsabilité applicable en cas d'accident pendant une expérimentation.

La loi Pacte du 22 mai 2019 exempte de responsabilité pénale les conducteurs de véhicules pendant les périodes où le système de délégation de conduite est activé.

C'est bien là la reconnaissance en creux d'une responsabilité du système robotique lorsque le système de délégation de conduite est activé.

En février 2020, la Commission européenne présentait ses stratégies en matière de données et d'intelligence artificielle, soulignant à cette occasion l'importance du développement d'une IA axée sur le facteur humain.

À cette occasion, Margrethe Vestager, vice-présidente exécutive pour une Europe adaptée à l'ère du numérique, déclarait : « Nous voulons que chaque citoyen, chaque travailleur, chaque entreprise ait une possibilité réelle de tirer parti des avantages de la numérisation. Qu'il s'agisse de conduire en toute sécurité ou en polluant moins grâce aux voitures connectées, ou même de sauver des vies à l'aide d'une imagerie médicale fondée sur l'intelligence artificielle, qui permet aux médecins de diagnostiquer des maladies plus précocement que jamais », (communiqué de presse du 19 février 2020).

C'est ce nouvel écosystème juridique, que l'Europe appelle de ses vœux, qui participe à maintenir l'intelligence artificielle au service du bien commun.

Mais au-delà, c'est aussi à une stratégie de régulation plus large, à l'horizon 2025, qu'appelle l'IA. Une stratégie qui devra être économique, technique, juridique et politique.

Quelle stratégie de régulation économique, technique, juridique et politique à l'horizon 2025 ?

C'est précisément l'intitulé d'une *masterclass* élaborée par Mines Paristech et l'Université PSL, en partenariat avec le cabinet Lexing Alain Bensoussan Avocats, qui sera organisée en 2021.

Cette formation initiée par le Professeur Arnaud de La Fortelle, directeur du Centre de Robotique, Mines ParisTech, visera à être capable, à l'issue de celle-ci, de :

- comprendre les objectifs et les stratégies de régulation engagées actuellement au niveau international (US, UE, Chine, France) et par les parties prenantes de grands secteurs industriels tels l'automobile, les technologies de sécurité, les transports publics ou encore l'industrie de la santé ;
- décrypter la politique de l'État sur l'IA et la robotique ;
- comprendre les défis scientifiques et technologiques auxquels l'IA et la robotique doivent répondre pour respecter les objectifs et les contraintes des régulations ;
- interroger la manière dont peuvent et doivent se fabriquer les politiques publiques en matière de réglementation de l'IA et des robots, confrontées aux innovations scientifiques, techniques et industrielles ;
- identifier le champ des possibles (solutions) en termes de régulation de l'IA et des robots au travers d'approches sectorielles croisées.

L'analyse de l'état de l'art et de la régulation engagée dans de grands secteurs industriels et de service, relatée par de grands témoins issus de chacun d'entre eux, permettra à cette occasion de nourrir et d'orienter le débat.

Conclusion

Dans le cadre de cette régulation que nous appelons de nos vœux, une évidence : dans l'attente d'un droit en devenir, les IA doivent être conçues selon l'« éthique *by design* ».

Cette orientation, gravée dans un code de l'honneur des codeurs, impliquera l'interdiction d'algorithmes liberticides et d'algorithmes indignes.

C'est la raison pour laquelle il devient urgent de mettre en place une charte universelle du code, dont le principe serait de refuser le code qui porterait atteinte aux droits de l'homme (concept de codage *Ethics by Design*).

Cette charte permettra de définir un droit à l'humanité dans les futurs algorithmes.

Le monde de demain devra être régulé par les droits humains de l'IA.

Intelligence artificielle et sécurité nationale

Par **Julien BARNU**
Conseiller industrie et numérique du secrétaire général de la défense et de la sécurité nationale (SGDSN)

Dans son rapport « Donner un sens à l'intelligence artificielle » de mars 2018, Cédric Villani désignait le secteur de la défense et de la sécurité nationale comme prioritaire en matière de développement et d'usage de l'intelligence artificielle en France, considérant que ce secteur pouvait « constituer un avantage comparatif de la France ». Dans cet article, je chercherai à mettre en lumière, de façon non exhaustive, les particularités de ce domaine pour l'usage de l'intelligence artificielle, en développant quelques exemples sur le fondement des missions du secrétariat général de la défense et de la sécurité nationale, en particulier dans les domaines de la cyberdéfense et du renseignement technique.

Apports et limites de l'IA : l'exemple de la cyberdéfense

L'IA est susceptible d'être une technologie clé dans le domaine de la cyberdéfense, mais ce domaine illustre également plusieurs de ses limites inhérentes. Ainsi, si l'on considère ce que peut apporter l'IA dans les différentes phases d'une attaque informatique :

- Dans la phase amont de protection des systèmes d'information, l'IA est susceptible de démultiplier les capacités de l'État et des entreprises à effectuer des diagnostics techniques, voire de mettre en place des mesures correctives, à bas coût, en automatisant la recherche de vulnérabilités et de défauts de configuration dans un réseau. De la même manière, l'IA permettra d'automatiser le travail d'évaluation de la sécurité des produits numériques. L'impact de ce progrès technologique sur le niveau global de cybersécurité est cependant discutable. D'une part, parce que ces mêmes outils d'IA, utilisés par des attaquants, leur permettront également d'identifier de façon massive et automatisée des failles dans les réseaux et dans les produits numériques, et, d'autre part, car, au-delà des failles techniques, ce sont les failles humaines qui constituent et continueront à constituer la principale porte d'entrée des attaquants dans les réseaux (introduction de clés USB piégées, ouverture de pièces jointes malveillantes, etc.). Par ailleurs, la généralisation de l'usage d'outils d'IA permettra vraisemblablement aux attaquants de forger du contenu piégé de plus en plus crédible, de façon automatisée, par exemple en parcourant les réseaux sociaux ou le contenu de messageries électroniques. Un tel usage de l'IA, permettant aux attaquants de mieux exploiter les failles humaines, pourrait même compenser son apport en matière de sécurisation technique. Le développement de l'IA dans le domaine cyber pourrait ainsi avoir un effet globalement nocif sur l'état général de cybersécurité du pays.
- C'est probablement dans la phase de détection des attaques que les technologies d'IA sont les plus prometteuses. En effet, les technologies de détection actuelles s'appuient majoritairement sur des « signatures », c'est-à-dire sur des éléments caractéristiques d'un attaquant (adresse IP d'un serveur d'attaque, code malveillant, etc.) connus et intégrés dans un dispositif de détection, comme une sonde ou un antivirus. Ces dispositifs ne peuvent donc pas, par nature, détecter des tentatives d'attaque nouvelles aux caractéristiques inconnues. L'intégration de l'IA, et en particulier des techniques d'apprentissage, dans les systèmes de détection, est en mesure de résoudre cette problématique, en permettant de détecter des anomalies dans un réseau, et donc des modes opératoires d'attaque jamais observés jusqu'alors. Cette application de l'IA à la détection des cyberattaques fait l'objet d'importants travaux, en France comme à l'étranger.

- En matière de réponse à une attaque informatique, l'apport de l'IA doit être relativisé pour plusieurs raisons. Tout d'abord parce qu'il est impossible de se reposer sur une analyse technique, fut-elle conduite par une IA, pour attribuer une attaque, c'est-à-dire pour en identifier l'auteur. Certes, l'IA permettra d'associer plus rapidement une attaque donnée à un « mode opératoire », c'est-à-dire à un ensemble d'outils connus et caractéristiques d'un groupe d'attaquants, mais un tel rapprochement ne suffit pas à attribuer une attaque. En effet, les outils d'attaque, dont beaucoup sont en vente sur Internet, peuvent aisément être réutilisés, et certains attaquants de haut niveau sont même connus pour introduire dans leurs outils de fausses preuves techniques afin de conduire à des erreurs d'attribution (en orientant vers un pays donné, par exemple). Là encore, l'IA renforcera vraisemblablement la capacité des attaquants à brouiller les pistes, en automatisant l'introduction de fausses preuves dans leurs outils d'attaque et en renforçant la crédibilité de ces dernières. Le volet humain du travail d'attribution, et en particulier le renseignement d'origine humaine, restera donc sans nul doute tout aussi crucial qu'aujourd'hui pour attribuer une attaque. En matière de réponse, il est encore plus difficile d'imaginer que l'IA puisse provoquer un effet de rupture, d'une part, parce que le choix de répondre à une attaque, qui suppose de l'avoir préalablement attribuée, demeurera un choix politique, et, d'autre part, parce que la nature de la réponse à une attaque restera évaluée au cas par cas, en fonction de nombreux facteurs, notamment géopolitiques, et ne se situe pas forcément dans le champ cyber : une réponse à une cyberattaque peut être diplomatique, politique, économique ou même militaire.

Un rôle de l'IA parfois mal compris : l'exemple du traitement de données hétérogènes

Un des défis majeurs des États dans le champ du renseignement est d'exploiter de façon optimale les importantes quantités de données collectées (données de connexion, images, captations audio, renseignement satellitaire, etc.).

Dans la poursuite de cet objectif, des briques d'intelligence artificielle peuvent naturellement compléter les outils d'analyse de données mis en œuvre, par exemple en dégageant des tendances ou en identifiant des singularités difficilement perceptibles par l'œil humain. Mais dans ce domaine, le rôle de l'IA, au moins à court terme, est souvent surestimé. La principale difficulté rencontrée aujourd'hui est en effet de disposer, avant même de songer à déployer des outils d'IA, d'une capacité souveraine à structurer les données collectées et à fournir aux analystes des outils numériques leur permettant de les exploiter efficacement.

Contrairement à une idée reçue, le succès des solutions logicielles, telles que celles fournies par le *leader* américain Palantir, ne repose pas sur l'IA, qui n'y est en réalité intégrée que très marginalement, mais sur la capacité de cette entreprise à agréger rapidement des données d'origines et de formats variés et à les présenter à l'analyste d'une façon qui lui permette de les visualiser, de les croiser et de les analyser selon son approche métier propre. La force de telles sociétés ne repose pas sur une quelconque avance technologique en matière d'IA, mais sur les développements informatiques qu'elles ont conduits en associant étroitement des clients d'horizons divers (services de renseignement, banques, industriels, etc.). Cette approche leur a permis de disposer d'une palette de plus en plus large de fonctionnalités rapidement déployables, qu'elles sont en outre en mesure d'adapter en permanence aux besoins de leurs nouveaux clients, dans une forme de cercle vertueux.

Le retard de l'Europe et de la France sur ce segment ne résulte donc pas de lacunes dans le domaine de l'IA. La difficulté des industriels français à proposer des produits compétitifs s'explique davantage par leur manque d'agilité et de culture commerciale, qui se manifeste par une incapacité

à s'approprier les besoins de leurs clients et à leur proposer, en travaillant étroitement avec eux, des offres adaptées à leur métier. Certes, des fonctionnalités fondées sur des technologies d'IA complèteront naturellement de tels outils de traitement de données, mais à court terme c'est moins de prouesses technologiques dans le domaine de l'IA que d'une meilleure approche client dont nous avons besoin pour voir émerger une offre nationale crédible.

Une spécificité du domaine de la sécurité nationale : la disponibilité de la donnée

Contrairement à d'autres domaines où le développement d'une IA souveraine se trouve freiné par la faible quantité de données disponibles, ces dernières étant largement aux mains de grandes entreprises technologiques étrangères, le domaine de la défense et de la sécurité nationale est particulièrement riche en données nationales : images captées par satellite, communications électroniques et métadonnées captées sur les réseaux télécoms, captations sonores ou vidéos, données informatiques collectées à des fins de cybersécurité, etc.

Mais, davantage que dans les autres domaines, il existe d'importantes contraintes en matière d'utilisation et d'ouverture de ces données : la protection du secret et le principe du « besoin d'en connaître » imposent en effet un cloisonnement strict de ces données. En particulier, l'accès aux données collectées par les services de renseignement sur le territoire national est très strictement encadré par la loi.

Constituer des jeux d'apprentissage largement accessibles à des fins de développement de technologies d'IA pour les besoins de la sécurité nationale constitue donc un défi de taille, qui nécessite une refonte profonde de la gouvernance de la donnée au sein des ministères régaliens, et une politique claire de gestion et de valorisation de la donnée, prévoyant des procédures de mise à disposition des données pour l'entraînement des algorithmes d'IA ou de test des algorithmes sur des données réelles.

Ce défi est d'autant plus grand que le développement de l'IA constitue un changement de paradigme pour la sphère de la défense : historiquement en avance sur l'innovation civile, elle est aujourd'hui contrainte de s'appuyer sur des acteurs du monde numérique qu'elle connaît relativement mal et avec lesquels elle doit imaginer de nouveaux modes de travail : mettre à disposition d'entreprises innovantes des lacs de données opérationnelles, en veillant à un contrôle strict des accès, s'appuyer sur des outils civils disponibles en source ouverte et les adapter, ou choisir d'investir dans la conception de systèmes d'IA spécifiques – quitte à ce que ces derniers soient moins performants, au moins temporairement, que des solutions commerciales étrangères.

La confiance dans l'IA : un enjeu crucial dans le domaine de la sécurité nationale

Les algorithmes d'IA produisent parfois des résultats aberrants pour la perception humaine et les systèmes d'IA peuvent échouer de manière inattendue. Selon les mots d'une ancienne directrice de la DARPA : "*The problem is that when they're wrong, they are wrong in ways that no human would ever be wrong.*" Les techniques d'apprentissage présentent par ailleurs un risque de biais, qui peut être involontaire (données d'apprentissage non représentatives) ou volontaire (données d'apprentissage modifiées par un tiers pour influencer le comportement du système). Ces techniques peuvent enfin présenter des résultats opaques, difficilement explicables – cette question bien connue de l'explicabilité de l'IA faisant l'objet d'importants travaux internationaux.

Ces limites de l'IA, qui posent la question de la confiance qu'on peut avoir dans un système d'IA, sont particulièrement problématiques en ce qui concerne l'usage de l'IA à des fins de défense et de sécurité nationale, et la rendent inadaptée à de nombreuses applications militaires. Il n'est pas possible à ce jour de dire si ces limites sont surmontables ou non.

Dans l'impossibilité de garantir avec certitude la confiance dans un système d'IA, nous serons amenés à fixer, pour chaque application de l'IA dans le champ de la défense et de la sécurité nationale, un « niveau de confiance » exigible. Cette approche nous impose toutefois de développer une capacité souveraine à évaluer un système d'IA, afin d'être en mesure d'estimer ce niveau de confiance, voire de le certifier – sur le modèle des certifications de sécurité délivrées par l'Agence nationale de la sécurité des systèmes d'information pour les solutions numériques traditionnelles. Or, la certification de l'IA, et en particulier des techniques d'apprentissage, est encore un objet de recherche, et le cadre et les modalités d'une telle certification restent à construire. Pour le développement des usages de l'IA dans le secteur de la défense et de la sécurité nationale, ce chantier constitue un défi majeur, et l'un des facteurs les plus limitants. Sans avancée substantielle dans ce domaine, l'IA restera cantonnée à des usages qui seront en mesure de faire gagner du temps à l'opérateur humain, mais non de créer un effet de rupture et de bouleverser les rapports de force.

Une IA ou des IA ? Représentations et relations avec les IA

Par **Arnaud de LA FORTELLE**
Centre de Robotique, MINES ParisTech – PSL

Cette simple question : « Une ou bien des intelligences artificielles ? » nous introduit dans la complexité de notre relation avec la notion même d'intelligence artificielle (IA). Comme pour l'homme, il n'y a bien entendu pas une seule intelligence et déjà la définition du concept se révèle difficile. Le fameux test de Turing – qui déciderait quand une intelligence artificielle serait du même niveau que l'homme – est de plus en plus contesté et ne donne aucune définition, au mieux il procède par la voie d'une comparaison et d'une analogie. Mais ce n'est pas tant le concept qui nous intéresse ici, que sa représentation qui influe sur notre usage de la technologie et notre relation avec celle-ci.

Dans un premier temps, nous aborderons les représentations, techniques ou artistiques, qui peuvent inconsciemment biaiser nos conceptions de ce qu'est, ou plutôt de ce que fait, l'intelligence artificielle. Dans un deuxième temps, nous confronterons ces représentations à la réalité des IA d'aujourd'hui. Il existe dans la manière dont nous construisons ces IA – comme nos autres outils, matériels et conceptuels – de nombreuses structures humaines qui façonnent les modalités dont les IA réagissent ou interagissent avec le monde. De plus, la réalité du déploiement des IA aujourd'hui est la pluralité, ce qui conduit à une vision substantiellement différente des représentations habituelles. Et ainsi se pose la question des interactions entre IA : pouvons-nous envisager ce que seraient des échanges entre plusieurs IA ? Et ce que cela signifierait pour les interactions entre humains et IA ? Finalement, ce raisonnement aboutit à se poser la question de toutes les interactions, non seulement entre ou avec les IA, mais aussi entre humains, avec des transformations probablement plus profondes encore qu'Internet et les réseaux sociaux, mais encore peu claires et où nous aurons certainement nos propres choix à faire.

Les représentations de l'intelligence artificielle

L'intelligence artificielle est assez systématiquement présentée au singulier. L'art propose de nombreuses images d'une telle IA, le plus souvent sans considération pour les humains (HAL dans le film *2001, Odyssée de l'espace*), voire leur étant ouvertement et agressivement hostile (Skynet dans le film *Terminator*). Parfois paradoxalement étouffante et protectrice (Colossus dans le film *Le Cerveau d'acier*). Le singulier engendre l'idée d'une concentration de puissance (omniscience et omnipotence) qui, au mieux, déséquilibre le monde. C'est une voie bien classique de construction d'un cadre narratif, mais le point est que ces représentations fictionnelles passent dans notre perception de la réalité, bien souvent inconsciemment. L'intelligence artificielle fascine par son omnipuissance.

Si l'on regarde la manière dont le domaine a évolué depuis des décennies, il est en effet frappant de lire les attributs que nous avons prêtés à nos programmes : langages, calcul formel, systèmes experts, apprentissage profond. Les mots sont importants dans une relation, et ceux que nous choisissons pour décrire nos avancées en la matière sont révélateurs à la fois d'attentes et de craintes très fortes à leur sujet. Pourtant, la réalité est un peu différente et les scientifiques œuvrant dans ces domaines n'ont d'ailleurs jamais eu la naïveté de croire ce que leurs propres

mots sous-entendaient. L'apprentissage profond (*deep learning*) induit l'idée d'une profonde connaissance quand il ne fait référence qu'à la profondeur d'un réseau de neurones artificiels, le nombre de ses couches, ce qui est une esquisse plutôt simpliste de cerveau ; et l'apprentissage consiste à optimiser les paramètres de différentes fonctions (formant un neurone artificiel) pour aligner les prédictions du réseau de neurones avec une (très grande) base de données. Même si l'on est justement impressionné par les avancées que ces techniques ont permises, on est très loin d'une vraie intelligence. Typiquement, ces programmes savent bien détecter des humains dans des images ou des vidéos, mais ne comprennent pas qu'une affiche (par exemple une publicité) ne peut pas être un humain, ce qui peut poser un problème quand on fait appel aux IA pour conduire des voitures. Ce qu'il faut retenir ici, c'est que la technique elle-même soutient, par son langage, une représentation qui met en valeur la puissance de l'intelligence artificielle.

Nous voudrions également souligner deux autres attributs importants de l'intelligence artificielle d'aujourd'hui : les données et la puissance de calcul (admirons encore ici l'utilisation du mot « puissance »). L'apprentissage artificiel (*machine learning*) tire son efficacité des gigantesques bases de données (des millions de livres, des milliards d'images) qui lui permettent d'optimiser ses réponses lors d'une phase d'optimisation (l'apprentissage) extrêmement gourmande en temps et en ressources. On peut donc avoir le sentiment que cet apprentissage conduirait à privilégier quelques rares acteurs possédant toutes les données et capables d'entrainer une IA qui serait omnisciente. De même, le besoin de puissances extrêmes de calcul abonde dans ce sens. Outre l'association que nous pouvons en faire avec le thème du *Big Brother*, il nous dirige encore une fois vers une conception au singulier de l'intelligence artificielle.

Réalité : structures et interactions

Tranchant résolument avec les paragraphes précédents, considérons maintenant les intelligences artificielles déployées aujourd'hui, ou plutôt les produits et services qui reposent sur ces techniques. Les premiers qui viennent à l'esprit sont les *chatbots* qui ont été largement disséminés par les pionniers mondiaux : Siri, Google Voice, Alexa et les autres. Ce sont effectivement des services centralisés, reposant sur un profilage des utilisateurs, qui ont à la fois la capacité de traiter des questions en langage naturel (un problème notoirement compliqué pour lequel les IA ont permis de gros progrès) et de gérer des associations d'idées afin de produire des réponses pertinentes à un utilisateur donné. Ces services sont généralement également associés à certains produits selon un modèle économique désormais bien maîtrisé. Mais il existe de nombreuses variations de ces services (banques, cinémas…) qui ne reposent pas toutes sur ces acteurs. Ensuite, il existe des services que l'on n'oserait pas nommer des intelligences, mais qui reposent sur les mêmes techniques, comme pour la prise de photos : on parlerait plutôt de filtre, mais c'est bien un réseau de neurones qui apprend à transformer des images brutes en des images qui nous plaisent. Ici, notons qu'un tel réseau se trouve dans chaque *smartphone* récent : on peut commencer à parler de pluralité des IA.

On voit ainsi apparaître un déploiement répandu des techniques d'IA pour proposer aux utilisateurs des services agrégés et adaptés à leurs besoins. La clef est donc bien entendu d'avoir accès aux désirs des utilisateurs, et aux données qui permettent de les prédire (d'où le succès des modèles d'affaires liés à la publicité prédictive). Suivant le chemin pris par ces données, on aura des modèles singuliers, si l'on centralise toutes ces données, ou des modèles pluriels si ces dernières restent locales, par exemple sur les *smartphones.* Il existe des projets visant à déployer de telles IA sur les *smartphones,* puisqu'ils en ont les capacités aujourd'hui : on aurait donc un traitement de nos données personnelles « chez nous » avec un accès aux données externes (achats, réservations, questions…) sur Internet *via* cette IA. Un modèle plutôt sympathique qui pêche essentiellement aujourd'hui par son manque de modèle d'affaires en comparaison des services

poussés par les géants du web. Mais on voit que la pluralité des IA est une réalité qui a des conséquences intéressantes.

Outre la pluralité de la localisation des données et des calculs, nous assistons à la spécialisation des IA en fonction de nos propres attentes. Qu'est-ce que j'attends d'une IA à ma disposition ? C'est une question qui obtient des réponses très variées, depuis le planificateur qui organise mes transports, mes finances ou mon travail de façon à profiter au mieux de mon temps, jusqu'au *coach* qui me stimule et me conforte. Au fur et à mesure que les IA déployées aujourd'hui se perfectionneront, nous devrions donc les voir apprendre et s'adapter à des situations très différentes. Or il est bien clair que cette faculté d'apprentissage est au cœur des capacités des IA : on aura alors autant d'IA que d'humains (peut-être même davantage), et aussi de multiples IA pour toutes sortes d'activités (cinéma, restaurant, voiture, bus, robots…). Chaque IA devra se développer de façon différenciée, et les données qui auront conduit à cette évolution ne seront pas partagées. Les IA seront nombreuses et finalement très hétérogènes.

En dépit de cette diversité, on peut imaginer que les techniques sous-jacentes resteront relativement communes. Les réseaux de neurones sont organisés en une architecture (couches, liens entre couches…) qui est souvent très liée aux objectifs que l'on souhaite atteindre. Le cerveau est un organe bien plus complexe et dynamique que tout ce qu'on peut imaginer aujourd'hui avec des réseaux de neurones artificiels, mais passons ces considérations et admettons des capacités décuplées pour les IA. Nous l'avons vu, il existe, dans la construction des IA aussi bien que dans leur apprentissage (les données), des structures humaines, implicites ou explicites, par exemple le langage ou la mémoire de nos demandes et actions. Ces structures devraient transparaitre dans les interactions (données demandées ou produites par les IA). Et puisque toute l'évolution initiale des IA sera d'optimiser leur réponse à nos questions, on peut imaginer qu'il ne devrait pas y avoir une trop grande distance entre les interactions que les IA pourront avoir avec nous ou entre elles.

Relations entre IA

On sait qu'il y a des liens très forts entre le langage, les structures de notre pensée et l'intelligence en général. Il est sans doute un peu présomptueux d'imaginer que ceci s'applique à des machines sans adaptation. En revanche, de ce que nous avons présenté, il semble réaliste que demain de nombreuses IA interagissent : on pourrait sous-traiter l'organisation d'un anniversaire au restaurant, avec choix de date et de menu, ce qui requiert de multiples aller-retours entre des IA qui organiseraient les activités et préférences de chaque entité, invités ou restaurateurs. La façon dont ceci se mettrait en place reste obscure, mais ce sera probablement un mélange de langage naturel (puisqu'il faudra bien que les IA le maitrisent) et d'échanges de données. La seule certitude, c'est que de nombreuses IA, diverses, devront interagir.

Ce que feront les IA de leurs interactions est difficile à évaluer. Le scénario optimiste est que cela conduise à une coopération ; le scénario pessimiste est que cette interaction produise, et ceci encore plus vite que chez les humains, des structures de domination entre IA. La réalité devrait se situer entre les deux, suivant la volonté et l'influence que nous aurons. Les limites sont extrêmement difficiles à placer, car la notion de nombre – qui est capitale dans les interactions – est certainement différente pour une IA et pour un humain. Avec combien d'autres une IA pourrait-elle se mettre en relation ? En quoi ceci changera la structure même des relations IA-IA ? La réponse est très ouverte. Mais cette structure interne au monde des IA aura évidemment un impact sur la relation des IA avec les hommes.

Conclusion

Cette notion d'évolution dans le monde des IA, qui semble inéluctable, a conduit à deux théories relativement différentes. L'une, portée par le transhumanisme, est l'idée de singularité (*The Technological Singularity*) : le processus s'emballera et à un moment – unique et sans retour – les deux mondes se déconnecteront (les adeptes de cette théorie en font la victoire de l'IA, mais on peut en faire d'autres lectures). La seconde est la coévolution : les humains seront entrainés dans la symbiose avec les IA, et les deux « vivront » ensemble. Dans les deux cas, il me semble que l'on retourne ici à nos représentations fictionnelles. S'il est certain que les outils et les concepts que l'homme a forgés ont contribué à le changer assez profondément, on peut en conclure que les IA, par une interaction bien plus forte avec nous qu'un marteau ou même qu'un réseau social, devraient également nous transformer sérieusement. Mais en faire une évolution ultime, comme une apocalypse ou un avènement rédempteur, parait davantage tenir de la foi que de l'analyse : beaucoup ont prédit la venue du surhomme, du fait de l'évolution ou des révolutions, mais il me semble qu'il n'est pas encore parmi nous, n'en déplaise à certains.

La conclusion que nous souhaitons donner de cette analyse, c'est qu'il y a un vrai enjeu à considérer la pluralité des intelligences humaines et artificielles. C'est une voie que nous avons déjà commencée à emprunter, même si nous n'en sommes qu'aux balbutiements des IA. Sans prêter aux IA de demain plus que nous ne devons, reconnaissons que c'est un miroir pour nous-mêmes. Il s'agit bien d'une évolution, que ce soit à travers l'usage des IA, usage qui pourrait se transformer en relation forte (regardons ce qui se passe avec les réseaux sociaux), ou à travers la société qui pourrait réguler ou détourner ces usages. Et comme dans nos sociétés nous souhaitons éviter la concentration de tous les pouvoirs en une seule entité, réfléchissons à distribuer convenablement les rôles des IA et apprenons à développer des relations saines, sans être captifs de représentations biaisées.

Intelligence artificielle et travail : le défi organisationnel

Par **Salima Benhamou** (PhD)
Économiste à France Stratégie au département Travail - Emploi et Compétences.

Introduction

L'intelligence artificielle occupe une place privilégiée dans les débats actuels sur l'avenir du travail (Benhamou et Janin, 2018) et les enjeux économiques et sociaux qu'elle soulève (Cécile, Lorenz et Saglietto, 2020 ; Benhamou, 2018). Depuis une dizaine d'années, cette technologie a fait des progrès considérables, grâce notamment à la collecte massive de données (les *big data*), à l'augmentation des capacités de calcul et au progrès algorithmique (Chollet, 2018). Ces progrès trouvent aujourd'hui de nombreux domaines d'applications et concernent de nombreux secteurs d'activité, publics comme privés (transports, santé, banques et assurances, logistique, la défense et la sécurité pour ne citer que quelques exemples), (Villani, 2018). Dans un avenir proche, les progrès technologiques pourraient bien permettre à l'IA d'effectuer des tâches de plus en plus complexes, comme celles qui font appel au raisonnement et à la prise de décision, et de rivaliser ainsi encore davantage avec les capacités cognitives humaines. La victoire de la machine au jeu de Go, les premiers véhicules autonomes ou encore les performances des logiciels d'aide au diagnostic médical sont autant d'exemples emblématiques des progrès accomplis dans le domaine de l'IA.

Certains observateurs voient l'IA comme une opportunité économique, grâce notamment à une meilleure optimisation des processus de production et à une baisse des coûts due à l'automatisation des opérations. D'autres, au contraire, la considèrent comme une véritable menace pour l'emploi, avec la disparition programmée de pans entiers d'activité, menaçant ainsi de nombreux métiers, peu qualifiés mais aussi hautement qualifiés (juristes, auditeurs, médecins, etc.). Entre ces deux scénarios extrêmes, aucun consensus ne se dégage aujourd'hui. Certaines études ont avancé l'hypothèse d'une automatisation massive des tâches existantes par les nouvelles technologies numériques, dont l'IA, pouvant mener à la destruction de près de la moitié des emplois au cours des deux prochaines décennies, aux États-Unis (Frey et Osborne, 2013) comme dans d'autres pays industrialisés (Bowles, 2014). Cependant, d'autres études (Arntz, Gregory et Zierahn, 2016) aboutissent à des chiffres beaucoup plus faibles, allant de 10 à 15 % des métiers automatisables. Si aucun consensus ne semble se dégager, cela tient en grande partie aux limites méthodologiques de ces études et à l'absence d'enquêtes statistiques à grande échelle sur l'IA, et en particulier sur le *machine learning* (Cécile, Lorenz et Saglietto, 2020).

Bien qu'il soit difficile de savoir aujourd'hui combien d'emplois seront détruits ou créés par l'IA, on peut toutefois avancer avec un certain degré de certitude que l'IA, comme toute autre technologie, transformera en profondeur la manière de travailler et même le contenu du travail lui-même, mais avec des impacts différenciés selon les secteurs d'activité (Benhamou et Janin, 2018 ; Benhamou, 2020). On peut ainsi s'attendre à des effets importants sur les compétences, sur la dynamique d'apprentissage individuelle et collective au sein des organisations comme sur la mobilité sur le marché du travail. Nous visons ainsi dans cet article à identifier les principaux enjeux que posera le déploiement massif de l'IA dans le travail, en nous appuyant notamment sur quelques illustrations sectorielles et en prenant en compte les potentialités de l'IA comme ses limites.

Les limites de l'IA

Les progrès de l'IA portent principalement sur le raisonnement logique, la représentation des connaissances, la perception ou le traitement du langage naturel. Toutefois, il ne s'agit pas de la pensée pour autant. Sans entrer dans les détails de ces technologies, il convient de retenir qu'il s'agit de technologies reproduisant une classification existante et répondant à un objectif bien défini, comme gagner à un jeu, identifier une pathologie particulière (comme une tumeur cancéreuse) ou encore « conduire » une voiture autonome selon des conditions de circulation spécifiques (conduire le jour sur autoroute, par exemple).

L'IA n'intègre qu'imparfaitement la complexité des phénomènes

Toutefois, si l'IA est capable d'effectuer des tâches simples mais aussi des tâches compliquées avec une réelle efficacité comparable à celle d'un humain, voire supérieure, l'exécution de ces tâches reste toujours basée sur des règles prédéterminées et relève de processus hautement standardisés, à partir de données massives codifiables. La principale limite de l'IA est de ce fait de ne pas pouvoir « dévier » des normes ou de penser par elle-même (Benhamou, 2020). Il s'agit donc d'une technologie déterministe et contrôlée, au sens où le programmateur de l'IA choisit l'architecture logicielle qu'il veut utiliser (type de réseaux de neurones, nombre de couches, etc.), la méthode d'apprentissage (algorithme d'initialisation et de mise à jour des poids associés à chaque neurone) et les données d'apprentissage utilisées.

Nous sommes par conséquent loin d'un dispositif doué d'une conscience de soi-même et doté d'une grande autonomie, échappant ainsi à son concepteur. Comme l'a souligné Benhamou (2020) à partir de plusieurs exemples sectoriels, cette limite rend difficile pour une IA la résolution des problèmes complexes : par exemple, la gestion de comportements humains imprévisibles, l'exécution de plusieurs tâches complexes en même temps, déterminer et analyser un lien causal entre plusieurs facteurs, ou encore faire preuve d'empathie et d'écoute envers les personnes en prenant en compte toute leur complexité (économique, sociale, humaine, psychique, etc.)… Ce n'est pas un hasard si les plus grandes réussites de l'IA fonctionnent principalement sur des images et des données, qui sont parmi les plus standardisées en termes de contenu numérique et donc déjà bien connues. Enfin, même si l'efficacité de l'IA est basée sur la disponibilité d'un grand nombre d'événements (souvent plusieurs milliers) et sur une puissance de calcul importante pour l'apprentissage, les résultats ne sont pas très généralisables d'une situation à une autre, ce qui constitue une autre limite importante de l'IA. La volumétrie des données ne garantit pas mécaniquement un haut niveau de qualité d'analyse et d'une prise de décision optimale à chaque situation ou évènement imprévisible. Ainsi, l'immense majorité des tâches de compréhension et de décision réalisées par les humains restent hors de portée pour les systèmes actuels à base d'IA. Aucun système expert « intelligent » utilisé dans le secteur de la santé n'est capable de prendre en charge des patients de manière totalement autonome, que ce soit dans le domaine du diagnostic, de la proposition thérapeutique, ou dans le domaine de la prévention des comportements à risque. De la même manière qu'un véhicule autonome n'est aujourd'hui capable d'anticiper une situation de conduite « non apprise ».

L'IA n'est pas une technologie autonome capable de penser par elle-même

Ainsi, la représentation que le grand public peut avoir de l'IA, à savoir une machine aussi intelligente qu'un humain, qui a conscience d'elle-même et qui peut faire des choix en toute autonomie, est très loin de la réalité. L'IA est présente dans nos *smartphones* pour gérer l'assistant vocal et utilisée pour optimiser l'affichage de pages de publicité personnalisées, mais les dispositifs existants ne sont pas près d'être dotés d'une conscience. Un système expert en santé peut identifier des tumeurs cancéreuses particulières, mais ne peut pas prendre en charge des patients complexes,

qui présentent plusieurs pathologies en même temps. Les progrès sont encore loin de laisser présager l'avènement d'une IA dite « forte », qui serait en fait comparable à l'intelligence humaine, en particulier dans sa capacité à comprendre le contexte et à faire appel au « bon sens », ainsi que dans sa capacité permanente d'apprentissage. Une telle réalisation semble bien hors de portée à l'heure actuelle, comme l'a souligné le chercheur Yann LeCun [1].

Malgré ces limites importantes, le caractère générique des technologies développées laisse entrevoir un impact sur l'ensemble des secteurs de l'économie. Tout l'enjeu est de pouvoir identifier les tâches automatisables de celles qui ne le sont pas, à tout le moins pas avant de nombreuses décennies, et quelles sont les conditions organisationnelles qui favoriseraient la complémentarité plutôt que la substitution humain-machine. En pratique, c'est la manière dont seront déployés les dispositifs, partagés les gains de productivité permis par l'IA, et les choix effectués en termes d'organisation des tâches et des équipes qui détermineront en grande partie de quel côté penchera la balance.

Substitution ou complémentarité humain/machine ?

Tous les métiers ne sont pas composés que d'une seule tâche mais de plusieurs, dont certaines ne sont pas automatisables. Certaines sont de nature « périphérique » et à faible valeur ajoutée et d'autres constituent le « cœur » de métier à forte valeur ajoutée. Selon le type de tâche, l'éventail des possibles peut aller de la simple suppression à la transformation, voire même à la création de nouvelles tâches au sein d'un même métier. L'IA pourra supprimer certaines tâches, si elles sont prises en charge totalement avec une meilleure efficacité et performance économiques, en apportant par exemple une qualité supérieure de précision dans l'exécution de certaines tâches, et ce à moindre coût. Mais d'autres tâches ne pourront pas être prises en charge par la machine.

Les tâches qui composent un métier ne sont pas toutes automatisables

Le rapport de France Stratégie (Benhamou et Janin, 2018) montre à cet égard, à partir de plusieurs illustrations sectorielles, que tous les métiers qui tirent leur force de leurs activités humaines et sociales et qui mobilisent des compétences cognitives faisant appel à la créativité et à la résolution de problèmes complexes seront préservés. Dans le domaine sanitaire et médico-social, les médecins généralistes et spécialisés comme les infirmières ou les aides-soignantes sont concernés. Dans le secteur des transports, l'activité de conduite en convoi sur autoroute peut disparaître à long terme, en raison du développement du véhicule autonome laissant présager une diminution probable du nombre de chauffeurs routiers à terme. Mais d'autres tâches de supervision peuvent également apparaître, pouvant conduire à transformer certaines professions dans ce secteur et mener à la création de nouveaux métiers destinés à la supervision de la gestion des flottes de véhicules, à la maintenance prédictive ou encore à l'accueil et la sécurité des véhicules. Ce phénomène n'est pas nouveau : la robotisation dans l'industrie automobile est un phénomène ancien qui a conduit au repositionnement des travailleurs sur des tâches de supervision. Tout l'enjeu sera d'identifier et d'anticiper les nouveaux besoins en compétences face à la transformation des métiers liée au déploiement de l'IA. De la même manière, la mise en place d'outils permettant de trier et de répondre aux requêtes les plus fréquentes et d'outils avancés de recommandations personnalisées entrainera une diminution du nombre d'employés et une augmentation de la complexité des tâches restant à traiter, que la machine ne pourra pas prendre en charge. Dans le secteur bancaire par exemple, le rôle des conseillers pourrait alors être renforcé et réorienté vers l'accompagnement

(1) « Tant que le problème de l'apprentissage non supervisé ne sera pas résolu, nous n'aurons pas de machine vraiment intelligente. C'est une question fondamentale scientifique et mathématique, pas une question de technologie. Résoudre ce problème pourra prendre de nombreuses années ou plusieurs décennies. En vérité, nous n'en savons rien », (Yann LeCun, Chaire Informatique et sciences numériques, 2016-2017, cours au Collège de France).

individualisé des clients, où le maintien d'une partie des interactions physiques sera nécessaire pour les cas les plus complexes et pour les populations fragiles.

Dans de nombreuses situations, l'IA est complémentaire à l'intervention humaine

Dans de nombreux cas, les dispositifs d'appareils basés sur l'IA sont déjà utilisés de manière complémentaire aux tâches effectuées par les humains, comme celles concernant l'aide à la décision. Ici, la tâche humaine n'est pas modifiée sur le plan conceptuel, mais le travailleur peut s'appuyer sur des systèmes capables de contribuer à améliorer les performances : diagnostic et recommandations thérapeutiques, service à la clientèle dans le secteur bancaire, etc. L'intervention humaine n'est nécessaire que pour des raisons de limitations technologiques ou d'« acceptabilité ». Comme expliquée précédemment, l'intervention humaine s'explique en partie par des limites inhérentes à la technologie elle-même. Par exemple, pour des situations dans lesquelles la technologie n'est pas mûre et ne semble pas près de l'être, en raison de la forte complexité de certaines activités, comme la conduite autonome en toutes conditions de circulation (la nuit, zones sombres, par exemple) ou la détection de polypathologies d'un patient qui relève d'une plus grande complexité d'analyse, de recueil et de traitement de données et d'une prise en charge pluridisciplinaire. L'intervention humaine peut aussi s'expliquer pour des raisons d'acceptabilité sociale, comme les annonces de diagnostic à un patient ou la prise de décision ayant des conséquences individuelles, interdite par la loi relative à l'informatique, aux fichiers et aux libertés. Les contacts humains dans les domaines nécessitant une interaction sociale sont souvent indispensables et ne peuvent donc pas être remplacés par l'IA.

En somme, toutes les activités où le degré de complexité liée à la prise de décision est trop élevé et pour lesquelles leur « force » provient du contact humain et des interactions sociales ou fondamentalement sociales, comme les activités liées au dialogue ou à la négociation, resteront effectuées par des humains. En revanche, toutes les tâches qui présentent une forte régularité basée sur des règles prédéfinies, telles que l'organisation, la planification, le contrôle à travers l'identification de fraudes ou d'anomalies, la gestion de l'information (collecte de données et traitement…) ont de fortes chances d'être automatisées ou déclassées par l'IA. Les fonctions de soutien sont particulièrement concernées et traversent de nombreux secteurs, tels que le commerce de détail et certaines fonctions de *back-office* dans le secteur bancaire, les assurances, le *marketing,* ou encore les services juridiques.

Promouvoir les organisations du travail apprenantes pour favoriser la complémentarité humains/IA

Bien que cela n'entraîne pas nécessairement une préoccupation pour l'emploi, qui lui-même change avec le temps, l'IA souligne l'importance d'« apprendre à apprendre » en continu. Dans cette optique, les formes d'organisation du travail basées sur une logique d'apprentissage continu sont particulièrement adaptées aux défis posés par l'intégration de l'intelligence artificielle. Comme le montre la récente étude de Benhamou et Lorenz (2020), les organisations apprenantes sont basées sur l'utilisation de formes d'organisation du travail qui favorisent le développement des compétences transversales et soutiennent l'apprentissage continu des salariés. Les exemples sectoriels ont aussi montré que les compétences transversales – la capacité à communiquer avec les autres et à influencer les décisions, la capacité à transférer des compétences et des savoir-faire organisationnels, à dépasser les règles et les cadres prédéterminés, la capacité à gérer les aléas – prendront davantage d'importance avec les déploiements de l'IA. Dès lors, les organisations apprenantes qui valorisent en priorité ces compétences transversales et l'apprentissage continu seront un levier essentiel de la complémentarité entre les machines et les travailleurs. Selon

Benhamou et Lorenz (2020), l'un des enjeux importants posés par l'évènement de l'IA sera d'accompagner les organisations du secteur privé et public dans leurs projets de transformations organisationnelles et managériales, à travers un programme national en faveur des organisations apprenantes, à l'instar des pays d'Europe du Nord. Les défis organisationnels doivent aussi être envisagés de manière cohérente avec l'évolution du système d'éducation et de formation continue, dont les caractéristiques actuelles dans plusieurs pays européens, dont la France, ne sont pas propices à un changement de paradigme organisationnel, inspiré du modèle apprenant.

Bibliographie

Arntz M., Gregory T. & Zierahn U. (2016), "The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis", OECD Social, Employment and Migration Working Papers, No. 189, OECD Publishing, Paris.

Benhamou S. (2018), "The world of work in 2030: Four scenarios" *in* NEUFEIND M., O'REILLY J. & RANFT F. (Ed.) *Work in the Digital Age: Challenges of the Fourth Industrial Revolution,* Rowman & Littlefield Intl., London-New York.

Benhamou S. (2020), "Artificial Intelligence and the Future of Work", *Revue d'économie industrielle,* vol. 169, n°1, pp. 57-88.

Benhamou S. & Janin L. (2018), "Artificial Intelligence and Work", Report France Stratégie to the Minister of Labour and the Minister of State for the Digital Sector.

Benhamou S. & Lorenz E. (2020), « Les organisations du travail apprenantes : Enjeux et Défis pour la France », rapport de France Stratégie, avril.

Bowles J. (2014), "The Computerization of European Jobs", *Bruegel,* Brussels.

Cécile C., Lorenz E. & Saglietto L. (2020), "Exploring the economic and social impacts of Industry 4.0", *Revue d'économie industrielle,* vol. 169, n°1, pp. 11-35.

Chollet F. (2018), *Deep Learning with Python,* Manning Publications, Sister Island.

Frey C. B. & Osborne M. (2013), "The future of employment", Oxford Martin Programme on Technology and Employment, Oxford University.

Villani C. (2018), "AI for Humanity: French Strategy for AI", https://www.aiforhumanity.fr/en/

Le futur du travail en présence de formes artificielles d'intelligence

Par **Yves CASEAU**
Membre de l'Académie des technologies

Introduction

Cet article s'intéresse à l'essor technologique de l'automatisation – des robots à l'intelligence artificielle – et son impact sur les emplois. À la suite de l'étude "The Future of Employment" de Carl B. Frey et Michael A. Osborne (2013), il y a eu de nombreuses réactions, dont la plupart conservatrices et prudentes. Je vais développer ici une vision de l'évolution du travail dans laquelle l'homme est complémentaire de ces nouvelles formes automatisées de production et de création de valeur, mais qui s'inscrit néanmoins dans la continuité de l'analyse de Frey et Osborne. Les capacités actuelles d'automatisation et la prise en compte de ce qui est à venir à partir des technologies développées aujourd'hui dans les laboratoires nous invitent à repenser, comme le proposent Michael Ballé et Eivin Reke (2020), le rôle des humains dans l'organisation du travail. Cette nouvelle vision du travail modifie l'organisation de l'entreprise, de façon interne, mais également en tant que participante à un réseau. L'automatisation permise par l'intelligence artificielle (IA) accélère la transformation vers « l'iconomie » (Volle, 2014), c'est-à-dire l'organisation de l'économie dans le cadre d'une exploitation pleine et entière des bénéfices de la technologie de l'information.

Cet article est organisé comme suit. La première section s'intéresse à la question pressante de l'impact de l'intelligence artificielle sur les emplois. Au fur et à mesure que les technologies cognitives développées en laboratoire trouvent leur place sur les lieux de production et de fourniture de services, le paysage de l'emploi se modifie. La deuxième section propose une courte analyse des capacités d'automatisation permises par l'IA, telles que déployées aujourd'hui et également de façon prospective. La troisième section cherche à en tirer les conséquences pour les entreprises de demain, en revenant sur la question de la transformation des emplois (Stiegler, 2015). Les entreprises, vues comme des communautés de destin, unies par une finalité sociétale et profitant d'une efficacité interne transactionnelle et collaborative supérieure au marché ouvert, continueront à prospérer dans le monde complexe de demain, mais au sein d'écosystèmes qui vont significativement évoluer.

1. Intelligence artificielle, automatisation et destruction des emplois

Une révolution numérique et cognitive qui s'appuie sur des humains

Le développement de l'intelligence artificielle s'inscrit dans le progrès constant de l'automatisation des tâches. Ce progrès n'est pas régulier : il y a des périodes d'accélération rapide, comme le développement du *deep learning* depuis 10 ans, et des périodes d'absorption des possibilités des technologies. La route vers l'automatisation n'est pas simple. Les annonces célèbres de Foxcon qui voulait en 2014 remplacer ses 300 000 employés par un million de robots n'ont pas été réalisées. En revanche, l'automatisation des entrepôts d'Amazon avec des robots KIVA est une réalité, tout comme celle des usines avec « un très petit nombre d'humains », des usines qui semblent entièrement autonomes lorsqu'on les visite. Cette réalité contrastée s'explique par le fait que l'automatisation des tâches non répétitives reste complexe. L'automatisation des emplois peut commencer souvent plus facilement par des emplois d'experts plutôt que par ceux des

généralistes, comme l'expliquent Brynjolfsson et McAfee dans leur livre *The Second Machine Age* (2015) : "*The main lesson of 35 years of AI research is that the hard problems are easy and the easy problems are hard... As the new generation of intelligent devices appears, it will be the stock analysts and petrochemical engineers and parole board members who are in danger of being replaced by machines. The gardeners, receptionists, and cooks are secure in their jobs for decades to come.*"

L'intelligence artificielle est une boucle d'apprentissage avec des humains à l'intérieur. C'est une des idées les plus importantes du livre *Human + AI* (Daugherty et Wilson, 2018). Le rôle de l'humain est multiple, il s'agit à la fois d'organiser l'apprentissage de la machine, de participer à son apprentissage, et d'utiliser la valeur produite par des algorithmes. À côté des cas de systèmes complètement automatisés, la grande majorité des systèmes intelligents sont des systèmes d'assistance, d'aide à la décision. Tout comme pour le jeu d'échecs, le meilleur agent pour résoudre un problème ou conduire un processus est un « centaure », la combinaison de l'homme et de la machine. Le codéveloppement du « centaure » (du couple agent humain et assistant intelligent) est une aventure formidable de réinvention des métiers, et une course à la création de nouveaux avantages compétitifs. La dimension de boucle d'amplification induit des avantages de premier entrant qui s'auto-entretiennent. L'homme joue également un rôle fondamental pour collecter, classer et qualifier les données.

Une révolution numérique qui supprime plus d'emplois qu'elle n'en crée

Depuis la publication de l'étude *"*The Future of Employment*"*, qui a annoncé que 47 % des emplois seraient menacés aux US par l'automatisation, le débat est intense. D'un côté, il existe de nombreuses études qui arrivent à des résultats du même ordre de grandeur. D'un autre côté, on trouve des études plus nuancées et moins pessimistes, comme celle de l'OCDE ou celle de McKinsey. Un des arguments est qu'une partie des « tâches », et non pas des « emplois », sont touchées. Mais l'expérience montre que les entreprises ont acquis la capacité à redistribuer les tâches pour transformer les gains efficacement en réduction de coûts salariaux, en dehors d'une hypothèse de croissance. Je soutiens l'hypothèse de l'étude Frey-Osborne, parce que les arguments opposés reposent sur une analyse de ce qu'on peut attendre des progrès de l'intelligence artificielle, qui est trop conservatrice. Je vous renvoie au livre *The Second Machine Age* (McAfee et Brynjolfsson, 2015), pour vous convaincre qu'une nouvelle vague d'automatisation arrive à grand pas. Pour reprendre une de leurs citations : *"Computers and robots are acquiring our ordinary skills at an extraordinary rate."* Le fait que le monde change aussi vite sous nos yeux doit d'ailleurs nous conduire à beaucoup de prudence sur les études que je viens de citer. Comme le remarque Neil Jacobstein à la Singularity University en 2016, ces études s'appuient sur une continuité des types de tâches à effectuer (ce qui permet d'appliquer le peigne de l'analyse de la future capacité à automatiser), une sorte de *"everything being equal"*, qui est probablement valide sur une courte échelle (quelques années), mais beaucoup plus discutable sur quelques décennies.

Le nouveau paysage des emplois

Pour comprendre les conséquences du déploiement progressif de nouvelles formes d'intelligence dans les entreprises, je m'appuie sur l'analyse proposée dans l'article de Susan Lund, James Manyika et Sree Ramaswamy, intitulé "Preparing for a new era of work" (2012). Cet article propose de séparer les emplois en trois catégories : la production, les transactions et les interactions. Les deux premières catégories sont celles qui vont être massivement touchées par l'automatisation : les emplois de production seront – dans leur grande majorité – remplacés par des robots tandis que les emplois liés aux transactions vont être fortement réduits par l'utilisation de l'intelligence artificielle, même si cela prendra un peu plus de temps. Il reste, pour un temps plus long, le domaine des métiers d'interaction. Les emplois de demain seront en grande majorité caractérisés par les échanges entre humains dans des dimensions émotionnelles qui dépasseront le cadre de

l'automatisation, tout en profitant de ce nouvel « environnement intelligent ». Par exemple, le jardinier, le masseur ou le plombier de demain seront des métiers technologiques, collaboratifs et sociaux, dans le sens où l'environnement intelligent déchargera de certaines activités et permettra de se concentrer sur l'essentiel (par exemple, le sens et le plaisir du jardin).

Dans cet univers qui se dessine, tout ce qui s'automatise devient une commodité, la valeur perçue se trouve dans les émotions et les interactions. Ces métiers d'interaction de demain ne sont pas issus de nouveaux domaines à créer, mais pour leur grande majorité la continuité des métiers d'interaction d'aujourd'hui. Santé, bien-être, ordre public, éducation et distraction vont continuer à être les principaux fournisseurs de travail pour les décennies à venir. Le jardinier du futur utilise probablement des robots, mais il vend une « expérience », dans le sens où il raconte une histoire. Une part importante de ces emplois se trouve dans l'économie quaternaire, qui développe « des nouveaux services incorporant des biens, la mise à disposition temporaire de biens, de personnes, ou de combinaisons de biens et de personnes ». L'évolution vers l'économie quaternaire est fort logiquement liée, comme le souligne Michelle Debonneuil, aux progrès des TIC (technologies de l'information et de la communication) qui permettent d'apporter des services véritablement personnalisés sur le lieu précis où ils sont nécessaires, y compris dans la gestion des femmes et des hommes qui rendent ces services de façon courte et ponctuelle.

2. Penser et préparer le futur de l'intelligence artificielle

Ce que l'IA sait faire aujourd'hui

Avant de parler de façon prospective de ce que les nouvelles formes d'intelligences augmentées pourraient être dans les entreprises de demain, il faut prendre conscience de l'omniprésence de l'IA dans les outils numériques d'aujourd'hui, ce qui est souligné dans le rapport de l'Académie des technologies (ADT, 2018). Comme le remarque Peter Domingos, l'auteur du livre *The Master Algorithm* (Domingos, 2015), *"People worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world."* L'IA est partout dans les interfaces de nos *smartphones* où elle facilite nos choix et assiste nos interactions. Elle est essentielle au e-commerce et par conséquent l'objet de toutes les attentions des géants du *web*. L'IA est à la source de l'accélération de la productivité des processus, à commencer par la *supply chain*. Les géants du *web* tels qu'Amazon et Alibaba ne sont pas simplement des champions des sites de vente, ce sont en premier lieu des champions de la *supply chain* en utilisant l'ensemble des données pour connaitre leurs clients le mieux possible. Comme le soulignent Eric Schaeffer et David Sovie dans *Reinventing the Product* (2019), l'IA est au cœur du *digital manufacturing* et de l'industrie 4.0, où elle permet de réinventer les produits et les procédés de fabrication.

L'IA d'aujourd'hui permet de résoudre, ou d'assister la résolution, de problèmes précis, de façon étroite et spécialisée. En revanche, la « boite à outils » de l'IA (ADT, 2018) est vaste et permet donc d'adresser des problèmes variés. L'application de l'IA au traitement de l'information, par exemple pour recommander ou faciliter des recherches, est au cœur du développement des plateformes numériques (Caseau, 2020). Les progrès de l'IA se retrouvent dans des formes multiples d'automatisation des processus, tels que les robots de RPA (*robotic process automation*) ou les *chatbots* (assistants conversationnels). Nous avons déjà pris l'habitude d'utiliser notre *smartphone* comme « prothèse cognitive », mais cette capacité ne va faire que s'étendre.

Comprendre ce qui se prépare

Les progrès en IA sont dominés depuis 10 ans par le développement des réseaux neuronaux profonds. Ces progrès ont permis de « passer une barrière », celle de la reconnaissance des images, des sons, et de façon plus globale de formes. Le livre de Martin Ford, *Architects of Intelligence* (2018), permet de comprendre la percée fondamentale qui a été accomplie, et une partie

des questions difficiles qui restent ouvertes. Le cerveau semble mieux décrit par l'expression "*a society of minds*", due à Marvin Minsky, que par un unique « réseau neuronal ». De la même façon, une partie importante de la recherche en IA se concentre sur l'intégration de multiples formes d'algorithmes d'apprentissage. Les différents succès spectaculaires de DeepMind, de AlphaZero à AlphaFold, sont une illustration de la richesse des méthodes d'IA et des techniques d'hybridation ou d'intégration. Même si nous utilisons dès aujourd'hui les résultats du *deep learning* dans la reconnaissance d'images, de sons ou de texte, il est facile de prévoir que les conséquences de cette « révolution de la décade 2010 » vont s'étaler sur les deux décennies à venir, au fur et à mesure que les « briques élémentaires » rendues possibles par l'apprentissage profond trouveront leur place dans de nouvelles architectures plus complexes de systèmes intelligents.

Les progrès en traitement du langage naturel, qui s'illustrent par les performances des traducteurs ou les robots écrivains, peuvent sembler superficiels, car il s'agit davantage de transformation à partir d'exemples que de compréhension. C'est ce qui explique la faiblesse des *chatbots* d'aujourd'hui, mais la deuxième révolution, celle du progrès sur le traitement sémantique du langage, est en vue. Les outils de traitement par réseaux neuronaux profonds, parfaitement illustrés par GPT3, sont capables de faire des choses remarquables (recherche, traduction, résumé) sans comprendre le sens des textes qu'ils manipulent. Mais l'énorme quantité de données ingérées et l'extrême richesse et complexité des structures produites permettent de commencer à utiliser ces outils comme des briques élémentaires qui encapsulent une « forme de compréhension », dans une approche où l'accumulation massive de données devient un « succédané d'expérience ». Une fois réifié et structuré, l'ensemble de tout ce que GPT3 peut nous dire sur un concept, par exemple « un *chat* », devient un *ersatz* de compréhension du concept. Il est donc probable que cette première étape du traitement du langage par des réseaux neuronaux soit le début de la véritable révolution que représenteront, dans les années à venir, des assistants cognitifs capables de compréhension, des textes comme des concepts.

De façon moins visible, parce que cela touche les systèmes et non pas les interfaces, les nouvelles formes d'intelligence artificielle appliquées au traitement de données permettent aux entreprises de réinventer l'optimisation de leurs processus dans un environnement complexe et imprévisible. Le changement qui est en train de s'opérer avec l'utilisation de données massives et de l'application de nouveaux algorithmes d'analyse de données est qu'il devient possible d'optimiser sans prévoir (Caseau, 2020). La génération précédente des outils repose sur l'extraction de modèles qui représentent notre compréhension du monde (des processus et de l'environnement) sur lesquels nous appliquons notre effort cognitif. Lorsque le monde change, il faut reconstruire sa représentation et ses outils. La prochaine génération d'outils de pilotage adaptatif utilisera des approches « boites noires » (par exemple avec le *deep learning*) dans lesquelles les modèles ne sont plus explicites et sont des objets transitoires, mis à jour de façon continue.

3. Le futur des entreprises et des emplois

Le principe de l'entreprise est plus pertinent que jamais

Une idée fréquente dans la prospective du travail prône la déconstruction de l'entreprise au profit d'une nouvelle structure constituée de places de marchés et de travailleurs indépendants. Cette thèse s'appuie sur l'approche célèbre de Ronald Coase, autour des coûts de transaction. Puisque la technologie joue un rôle de fluidification (communication, échange, enregistrement, etc.) de ces transactions, plus la technologie progresse, moins il est nécessaire d'avoir une organisation dédiée. On passe de l'entreprise à l'entreprise 2.0, puis à l'entreprise étendue, puis au réseau d'agents. Je pense que cette analyse est fausse, que nous allons continuer à voir des entreprises (dans un sens traditionnel, pas seulement des marques) dans les décennies à venir, pour plusieurs raisons. Le monde numérique est une économie à coûts fixes, ce qui renforce la concentration et la création de monopoles. C'est très bien expliqué par les penseurs de « l'iconomie » (voir le livre de Michel

Volle, 2014). Le monde de demain est un monde de plus en plus complexe, et la complexité est défavorable à l'hyperspécialisation. La complexité augmente le coût de transaction : elle crée une « taxe de communication » qui est d'autant plus élevée que l'on décompose ce qu'on cherche à faire. La complexité est défavorable à l'abstraction et à la décomposition qui conduisent aux « places de marché ». Au contraire, la réaction des géants du *web* est à l'opposé : intégrer les compétences en équipes transverses pour minimiser les coûts de transaction. On pourrait même dire que les tâches qui se prêtent à la désintermédiation par des plateformes sont les premières candidates à l'automatisation, au fur et à mesure des progrès évoqués dans la section précédente. La complexité maximale que sait traiter une entreprise augmente à la vitesse permise par la baisse du coût élémentaire de transaction, rendue possible par la technologie.

L'autre utopie très à la mode est celle de l'entreprise ATAWAD ("*anytime, anywhere, any device*"). Cette vision cherche à nous affranchir du lieu et du temps : on peut travailler de n'importe où, au moment où on le souhaite et au rythme de son choix. Au contraire, la complexité impose le travail synchronisé en mode « liens forts », c'est-à-dire de façon colocalisée. Il ne s'agit pas d'une exigence absolue ou constante, comme le développement du travail à distance à la suite de l'épidémie du Covid-19 vient de le montrer, mais la majorité de la création de valeur aujourd'hui est faite par des équipes qui travaillent ensemble. Puisque cette approche de colocalisation n'est pas *scalable*, il y a forcément une organisation en réseau distribué, avec une importance sans cesse croissante des nouvelles technologies de communication et de collaboration.

Le paysage du réseau des entreprises de demain

Ma vision du paysage de l'emploi est un réseau multi-échelle, depuis les grandes entreprises multinationales d'aujourd'hui jusqu'aux autoentrepreneurs, dans lequel le mouvement de polarisation (consolidation des grandes plateformes et multiplication des « pico-entreprises ») se poursuit et s'amplifie. Nous aurons toujours des entreprises dans 20 ou 30 ans ; le phénomène d'ubérisation du travail n'aura pas dissous le concept de l'entreprise. La mondialisation et la numérisation conduisent à la concentration. En addition à l'argument précédent d'une économie de coûts fixes qui favorise l'économie d'échelle, les effets de réseaux – en particulier dans les marchés bifaces et dans le développement d'écosystèmes autour de plateformes – donnent un avantage important au plus gros joueur (souvent le premier mais pas forcément). Cette concentration des plateformes conduit également à la croissance des écosystèmes qui leur sont associés, ce qui peut créer des opportunités pour une multiplicité d'acteurs locaux. Prenons l'exemple des capacités d'intelligence artificielle développées et exposées par Google. Il est plus que probable que si ce développement est couronné de succès, il va contribuer à la croissance forte de Google. Mais il va également ouvrir des champs possibles à un rythme supérieur que ce que Google peut produire, ce qui signifie qu'une partie encore plus importante de valeur va apparaitre « à la frontière », lorsque d'autres acteurs vont utiliser ces technologies mises à disposition par Google pour résoudre d'autres problèmes que ceux qui l'intéressent. On voit la même chose avec la croissance d'iOS comme plateforme mobile : au fur et à mesure que les capacités sont ajoutées dans la plateforme de développement de l'iPhone – on pense ici bien sûr à Siri –, le domaine fonctionnel rendu possible à la communauté des applications mobiles augmente plus vite que ce qu'Apple en retire pour ses propres services.

L'idée que nous allons tous vivre de notre créativité tandis que les machines s'occuperont de la production est naïve et probablement fausse. Le tissu de multinationales a besoin de nouveaux talents, et en particulier de créatifs et de *designers*, mais dans un petit nombre par rapport aux emplois supprimés en conséquence de l'automatisation. En revanche, le monde « frontière » des opportunités de services, qu'il s'agisse d'adaptation au besoin d'une communauté ou d'un individu, a une structure beaucoup plus riche et étendue que l'on pourrait qualifier de « fractale » ou de « multi-échelle ». Dans ce monde de l'interaction, il existe des opportunités à différents niveaux

de talents, qui peuvent coexister. Le service d'interaction se déplace difficilement (en tout cas avec un coût) contrairement à une expérience digitale. Les « artisans de la personnalisation » de masse peuvent opérer sur des échelles géographiques différentes selon leur talent. Cette renaissance de « l'artisan de proximité » risque de plus de se trouver facilitée par la priorisation de ce qui est local sur ce qui est global, en contre-réaction à la mondialisation et en soutien de la lutte contre le réchauffement climatique.

Conclusion

Pour conclure, je vais souligner quatre idées qui me semblent importantes pour comprendre l'impact des nouvelles formes d'intelligence artificielle sur le fonctionnement des entreprises :

- L'intelligence artificielle n'est pas une technique isolée, c'est un outil, une modalité des solutions logicielles.
- L'intelligence artificielle est construite et est déployée dans une collaboration entre les hommes et les machines.
- L'intelligence artificielle est multiple, il existe une grande diversité de méthodes qui gagnent à être combinées, soit par hybridation, soit par assemblage dans un « système de systèmes ».
- L'intelligence artificielle est un « absorbeur de complexité » ; plus elle se développe, plus elle permettra aux entreprises de demain d'aborder et de développer des nouveaux territoires de création de valeur.

Bibliographie

ADT (2018), « Renouveau de l'Intelligence Artificielle et de l'Apprentissage », rapport de l'Académie des technologies.

BALLÉ M. & REKE E. (2020), "Do We Still Need People?", *LindedIn Pulse,* 10 août 2020.

CASEAU Y. (2020), *L'approche Lean de la transformation digitale – Du client au code et du code au client*, Dunod.

DAUGHERTY P. & WILSON H. J. (2018), *Human + Machine: Reimagining Work in the Age of AI,* Harvard Business Review Press.

DOMINGOS P. (2015), *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, Basic Books.

FORD M. (2018), *Architects of Intelligence: The Truth about AI from the people building it,* Packt Publishing.

FREY C. B. & OSBORNE M. (2013), "The Future of Employment", Oxford Martin Programme on Technology & Employment, working paper.

LUND S., MANYIKA J. & RAMASWAMY S. (2012), "Preparing for a new era of work", *McKinsey Quaterly*.

McAFEE A. & BRYNJOLFSSON E. (2015), *The Second Machine Age*, Norton.

SCHAEFFER E. & SOVIE D. (2019), *Reinventing the Product: How to Transform your Business and Create Value in the Digital Age*, Kogan Page.

STIEGLER B. (2015), *L'emploi est mort, vive le travail !*, Fayard - Mille et une nuits.

VOLLE M. (2014), *Iconomie,* Xerfi, Economica.

Algorithmes et droit pénal : quel avenir ?

Par **Elise BERLINSKI**
ESCP
Imane BELLO
Avocate à la Cour
Et **Arthur GAUDRON**
Chercheur au Centre de Robotique

« Les algorithmes brevetés inondent le système de justice criminelle. Les systèmes de *machine learning* déploient les agents de police dans les quartiers prétendus "chauds". Les laboratoires de police scientifique utilisent des logiciels probabilistes pour analyser les indices médico-légaux. Les juges utilisent des "instruments d'évaluation du risque" automatisés pour décider qui peut sortir sous caution, voire même quelle peine imposer » (Wexler, 2017).

Les algorithmes d'intelligence artificielle (IA), plus spécifiquement de *machine learning* (ML), sont à présent une réalité dans le droit pénal états-unien et ne sauraient tarder à le devenir en France, comme en témoignent par exemple les discussions sur leur application dans ce domaine du rapport de la CNIL (Demiaux & Si Abdallah, 2017). Dans cet article, nous nous intéressons aux implications de l'utilisation du ML dans le cadre du droit pénal français, et la manière dont celui-ci pourrait le modifier.

L'application d'algorithmes de ML dans le contexte du droit pénal vise à la création d'indices dans l'optique d'assister à la détermination du déroulement effectif d'un acte répréhensible, ou à l'orientation d'une enquête. L'utilisation de tels algorithmes demande de se poser des questions sur les données qui les alimentent, d'une part, et les calculs menés, d'autre part. Ces données proviennent de plus en plus fréquemment de réseaux sociaux, comme en témoigne le discours de Lord Gross (2018). Elles sont souvent comportementales, construites par désagrégation et réagrégation, et visent à reconstituer un sujet, généralement traité comme représentatif d'un sujet réel, donc capable d'en dévoiler l'essence (Chamayou, 2015), alors même que des études montrent qu'il n'existe pas de correspondance évidente entre ce sujet numérique construit et un sujet référent réel (Goriunova, 2019). Il nous semble donc urgent de comprendre le prisme de médiation (Hansen & Flyverbom, 2015) par lequel ces technologies représentent les personnes et leurs effets.

Après une rapide présentation du cadre juridique pénal en France, nous décrivons les méthodes de construction du sujet numérique, avant d'en tirer des conclusions sur le phénomène observé.

Les preuves et raisonnements juridiques en droit pénal

En droit pénal, une infraction est constituée par la réunion des éléments suivants : un élément légal (la répression par un texte de loi), un élément matériel (un comportement en lien de causalité avec le résultat entraîné) et un élément moral (la participation libre à un fait dont on connaît le caractère illégal). Les infractions peuvent être matérielles (le dommage est réalisé) ou immatérielles (l'accomplissement de l'acte incriminé seul suffit pour être qualifié d'infraction, c'est par exemple le cas de la fabrication de fausse monnaie). Par ailleurs, une tentative de crime est elle aussi répréhensible dès l'instant qu'« elle n'a été suspendue ou n'a manqué son effet qu'en raison de circonstances indépendantes de la volonté de son auteur » (code pénal - article 121-5, non daté). Il est donc envisageable que les officiers de police utilisent des algorithmes de ML de

manière à déterminer la probabilité qu'un crime ou délit ait lieu puis à déclencher une enquête en vue d'en déterminer son existence effective.

A titre d'exemple, l'administration fiscale (tant la Direction générale des finances publiques que la Direction générale des douanes et droits indirects) utilise des algorithmes de ML en vue de détecter des incohérences susceptibles de constituer des schémas de fraudes ou des signaux faibles de potentielles fraudes (Bello & Daoud, 2020).

Afin d'entrer en éventuelle condamnation, le juge doit analyser la caractérisation de l'infraction qui se situe dans le passé et se matérialise par un comportement. Cette analyse s'appuie sur des preuves dont le régime en droit pénal est dit « libre » : aux termes de l'article 427 du code pénal, « Hors les cas où la loi en dispose autrement, les infractions peuvent être établies par tout mode de preuve et le juge décide d'après son intime conviction. Le juge ne peut fonder sa décision que sur des preuves qui lui sont apportées au cours des débats et contradictoirement discutées devant lui. » Ainsi, le résultat de modélisations statistiques, de systèmes d'exploration de données algorithmiques, s'il ne peut pas, à proprement parler, servir comme élément de preuve de la réalisation d'un acte, peut toutefois influencer l'appréciation souveraine du juge.

Ainsi, bien qu'il n'existe pas de disposition légale réprimant un acte qui ne soit pas réalisé (de manière matérielle, immatérielle, ou sous forme de tentative), nous observons que les résultats des algorithmes peuvent influencer le juge (et sa perception du réel), d'une part, et, d'autre part, orienter l'enquête. Dans le premier cas, même si les résultats algorithmiques ne pourront être utilisés sans preuve tangible de la réalisation de l'acte, il est possible d'imaginer qu'ils puissent orienter la détermination de la peine et appuyer la conviction du juge. En effet, selon le principe d'individualisation des peines, la personnalité et la singularité d'une personne sont prises en compte dans le prononcé de sa peine. Dans le second cas, l'algorithme orientera le choix de réaliser une enquête et la façon dont celle-ci sera menée. Dans tous les cas, il nous semble nécessaire de comprendre le prisme pour penser le phénomène auquel nous faisons face.

Création du sujet numérique

Deux blocs constituent ce sujet : les données et les calculs algorithmiques qui visent à déterminer la probabilité qu'un individu ait commis une infraction ou non.

Les données décrivant le sujet auquel on s'intéresse peuvent provenir, d'une part, du domaine juridique (l'ensemble des éléments enregistrés liés à une affaire, pouvant s'étendre des déclarations à la police jusqu'au jugement), d'autre part, du domaine public/privé (collectées par un ensemble de capteurs numériques, dont les réseaux sociaux).

Ce profil singulier doit par ailleurs être « comparé » (par un modèle) à des bases de données d'observations, ou données d'entraînement, qui informent, d'une part, des infractions commises par une personne, d'autre part de leurs données comportementales collectées par divers capteurs numériques. À l'heure actuelle, les réseaux sociaux représentent de bonnes bases de données, puisqu'ils placent des capteurs sur l'ensemble du *web* de manière à constituer des profils comportementaux, qu'ils monétisent ensuite à des organisations qui veulent justement évaluer des profils de risque (crédit, assurance…) (Arvidsson, 2016 ; Fourcade & Healy, 2013).

L'algorithme de calcul n'est pas déterminé par une théorie (sociologique par exemple), mais se forme en apprenant à partir de la donnée et d'inférences statistiques. Nous nous concentrons sur les algorithmes supervisés, c'est-à-dire qui apprennent à partir de données labélisées (qui portent sur des personnes dont on sait si elles ont commis des infractions et, le cas échéant, lesquelles). Ce choix est motivé par le fait que ces algorithmes sont plus performants, et parmi les plus répandus (LeCun, 2016), et qu'il existe de la donnée pour les alimenter. Par ailleurs, nous nous intéressons

aux tâches de classification, qui donnent la probabilité qu'un élément appartienne à une certaine classe (par exemple la probabilité qu'une personne ait commis une infraction donnée – classe 1 – ou non – classe 0).

Prisme de perception : le sujet numérique effectif

Le sujet numérique prend deux formes singulières : une forme « diffuse », comme masse informe de données ; une forme « effective », opérationnalisée par un algorithme.

Le sujet numérique diffus ne constitue pas une information intelligible pour l'humain. Nous nous concentrons donc dans cette section sur le sujet numérique effectif, qui est le résultat d'un traitement algorithmique. Celui-ci se matérialise sous forme d'un score qui représente la probabilité qu'une personne commette telle infraction, étant donné la représentation numérique que l'on a de celle-ci et les données d'entraînement utilisées. L'algorithme se présente donc comme l'étape de distillation permettant d'obtenir l'essence de ce sujet diffus et dépend de deux types de données : celles récoltées sur la personne en particulier et celles qui ont servi à entraîner l'algorithme.

Il existe différents algorithmes de ML supervisés, et chacun de ces algorithmes transforme la donnée différemment, menant à des sujets effectifs différents à partir d'un sujet diffus unique. Nous proposons une première compréhension de ces différents sujets selon l'algorithme choisi, en nous concentrant sur trois d'entre eux particulièrement connus et dont les fonctionnements diffèrent fortement : la machine à vecteur de support (SVM), les arbres hiérarchiques et les réseaux de neurones artificiels (ANN pour *artificial neural network*).

- Le SVM est un algorithme de classification qui s'appuie sur deux grands principes. D'une part, il recherche une frontière telle que la distance significative des classes qui lui sont adjacentes soit maximale. D'autre part, comme les observations ne sont pas nécessairement séparables dans l'espace mathématique auxquelles elles appartiennent, cet algorithme permet de projeter celles-ci dans des espaces plus complexes dans lesquels la structure des données, c'est-à-dire l'existence de classes comme sous-espaces séparés, sera plus saillante. Ainsi, cet algorithme crée des sujets numériques effectifs dont les différentes opérations sur ceux-ci visent à les discriminer au maximum. Cependant, cette discrimination se fait dans un espace abstrait qui nous est inintelligible, au sein duquel certaines qualités du sujet sont alors « amplifiées ». Ces qualités se présentent cependant comme des abstractions mathématiques indépendantes de notre culture et inintelligibles à l'humain. La re-projection de ce résultat dans le monde physique pourrait mener à des discriminations difficilement explicables. Par exemple, sans préjuger de l'algorithme utilisé, l'utilisation d'un algorithme prédictif par un sheriff aux États-Unis a mené à des situations de harcèlement de personnes, sans raison apparente à nos yeux (Holmes, 2020).
- Les arbres hiérarchiques, simples ou complexifiés (forêts aléatoires, *boosting*) sont des algorithmes de classification, qui hiérarchisent les variables descriptives de la personne (donc ses données) et les découpent par régions de valeurs telles que les régions finales correspondent à des classes. Contrairement au SVM précédent, ces arbres attachent une grande importance aux variables constitutives du sujet numérique diffus. Celles-ci ne sont pas projetées, mais conservées et minutieusement découpées jusqu'à ce que l'on estime le résultat obtenu satisfaisant. Ces arbres supposent qu'il est possible de décrire par hiérarchisation de caractéristiques les individus, à travers leur sujet numérique effectif, et d'en tirer des conclusions quant à la nature plus ou moins criminelle de ces individus. Le logiciel COMPASS hiérarchisait aussi des dimensions (Angwin *et al.*, 2016), mais ici celles-ci sont créées à partir du sujet diffus, qui dicte donc les valeurs selon lesquelles un individu peut être classifié comme représentant ou non un danger, et un traitement algorithmique plus ou moins complexe. Nous proposons donc de nommer cette forme de sujet numérique « le sujet hiérarchisé ».

- Enfin, les derniers types d'algorithmes que nous abordons sont les réseaux de neurones artificiels (ANN). Ceux-ci ont connu ces dernières années un succès fulgurant, en particulier grâce à leurs performances impressionnantes dans les tâches de reconnaissance d'image. Ces algorithmes sont structurés par des « couches » : une couche d'entrée, à travers laquelle les données sont alimentées, une couche de sortie, qui correspond à la classification recherchée, et des couches intermédiaires. Chacune des couches intermédiaires a une fonction spécifique, attribuée automatiquement par l'algorithme, permettant de capturer par découpages successifs les différences topologiques existant entre les classes. Ces algorithmes, en particulier sous forme de réseaux profonds, ont tendance à mémoriser la forme des phénomènes plutôt qu'à chercher à en découvrir des structures explicatives (même dans un monde algorithmique), ce qui s'illustre par leur grand nombre de paramètres, par exemple le VGG19 est réputé pour ses très bonnes performances et contient 144 millions de paramètres. Ainsi, ils supposent qu'en appliquant un nombre de décompositions assez importantes, l'entité considérée sera presque « indifférenciable » d'une entité reconstituée (sous forme d'agrégat réticulaire de données de plusieurs entités), ce qui suppose une vision relativement déterministe et circonscrite du monde. Nous proposons donc de nommer ce troisième sujet effectif « le sujet circonscrit ». Par exemple, estimer qu'analyser une photo suffit à déterminer une personne, comme ce fut le cas pour Robert Williams, qui écopa d'une garde à vue de 30 heures (Le Monde avec l'AFP, 2020), ne représente-t-il pas une perception circonscrite ? Ce prisme de perception tend à ignorer la singularité du sujet, et soutient l'idée qu'il est toujours possible d'en déterminer une essence plus ou moins criminelle.

Sujet numérique diffus et effectif dans le droit pénal

Nous avons vu que les algorithmes de ML, alimentés par des données individuelles et comportementales, constituent des sujets numériques comme prismes d'observation des individus, pouvant servir d'indices à la construction future d'éléments de preuve, d'une part, ou pouvant servir à orienter l'enquête, d'autre part.

Dans un premier temps, ces sujets numériques, que nous nommons « diffus », sont constitués par une masse informe de données inintelligibles à l'humain, à cause de leur volume, structure ou mode d'agrégation. À ce stade, la matière d'analyse a été sélectionnée algorithmiquement et est aveugle à la question des dimensions conventionnelles ou culturelles d'analyse, qui devraient être, selon nous, privilégiées. Pour être clairs, ce sont les fournisseurs de données (réseaux sociaux, *data-brokers* ou autres) qui auront déterminé ce qui est à inclure ou non dans l'analyse, mais aussi les relations à prendre en compte dans ces données.

Dans un deuxième temps, lorsqu'un sujet « effectif » est construit, celui-ci se résume en la probabilité qu'il ait commis (selon l'ensemble des variantes possibles) un crime ou un délit. Ce sujet effectif est le résultat de diverses transformations qui impliquent que le sujet (réel) est observé à travers un prisme particulier, dont potentiellement personne n'est conscient. Nous avons caractérisé trois sujets effectifs spécifiques construits à partir d'algorithmes particuliers. Premièrement, le sujet « amplifié » par un algorithme de SVM, qui par abstraction mathématique dans des espaces complexes permet « d'amplifier » la discrimination entre divers sujets, amplification faite selon des règles inaccessibles à l'humain. Deuxièmement, le sujet « hiérarchisé » par des arbres hiérarchiques, où le sujet diffus segmenté et recombiné permet de déterminer les traits saillants d'un profil « criminel » ou non spécifique. Ces traits saillants, encore une fois, sont déterminés algorithmiquement, et rien ne permet de savoir si d'autres dimensions du sujet existent que l'algorithme n'estime pas importantes. Enfin, le sujet « circonscrit » par l'ANN, qui par zooms successifs tente d'en déterminer une essence déterminante. Cette essentialisation résorbe totalement le déterminisme de l'aléa (Longo, 2019).

Comprendre ces processus nous semble central, d'autant qu'ils pourraient modifier les frontières du droit pénal. En effet, jusqu'à présent, la responsabilité d'un individu face à la loi consiste en

la possibilité de déterminer avec assez de certitude si ce dernier a commis un acte délictueux ou criminel. Cela signifie, d'une part, que l'enquête porte sur une action et, d'autre part, que cette action se situe dans le passé. À l'inverse, l'utilisation du ML modifie le point de focalisation de l'action vers le sujet et du passé vers le futur, ou de la réalisation vers le potentiel criminel (puisque les données peuvent être ultérieures, ou être utilisées pour lier un comportement à un crime potentiel). Enfin, le sujet observé n'est pas le sujet physique, mais le sujet numérique. Ainsi, ces algorithmes incarnent une puissance de changement importante, qu'il convient de continuer d'explorer de manière à être conscients de leurs effets potentiels et à mettre en place les régulations nécessaires.

Bibliographie

ANGWIN J., LARSON J., MATTU S. & KIRCHNER L. (2016), "Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks", *ProPublica*.

ARVIDSSON A. (2016), "Facebook and Finance: On the Social Logic of the Derivative", *Theory, Culture & Society*, 33(6), 3–23. https://doi.org/10.1177/0263276416658104

BELLO, I. & DAOUD E. (2020), "Les nouveaux moyens de lutte contre la fraude fiscale", *Revue Lamy Droit des Affaires*.

CHAMAYOU G. (2015), « Avant-propos sur les sociétés de ciblage », *Jef Klak*, 2, 1–12.

Code pénal — « Article 121-5 ».

CORNU G. (2020), *Vocabulaire juridique* (13e éd.) , PUF.

DEMIAUX V. & SI ABDALLAH Y. (2017), « Comment permettre à l'homme de garder la main ? », CNIL.

FOURCADE M. & HEALY K. (2013), "Classification situations: Life-chances in the neoliberal era", *Accounting, Organizations and Society*, 38(8), 559–572. https://doi.org/10.1016/j.aos.2013.11.002

GORIUNOVA O. (2019), "The Digital Subject: People as Data as Persons", *Theory, Culture & Society*, 36(6), 125–145. https://doi.org/10.1177/0263276419840409

GROSS J. (2018), "Speech By Lord Justice Gross Disclosure – Again", *Judiciary of England and Wales*.

HANSEN H. K., & FLYVERBOM M. (2015), "The politics of transparency and the calibration of knowledge in the digital age", *Organization*, 22(6), 872–889. https://doi.org/10.1177/1350508414522315

HOLMES A. (2020), "A sheriff launched an algorithm to predict who might commit a crime. Dozens of people said they were harassed by deputies for no reason", Business Insider. https://www.businessinsider.fr/us/predictive-policing-algorithm-monitors-harasses-families-report-2020-9

Le Monde avec l'AFP (2020), « États-Unis : Un Américain noir arrêté à tort à cause de la technologie de reconnaissance faciale », *Le Monde*.

LECUN Y. (2016), « Les Enjeux de la Recherche en Intelligence Artificielle. Leçon Inaugurale. Chaire Informatique et sciences numériques », Collège de France.

LONGO G. (2019), "Letter to Turing", *Theory, Culture & Society*, 36(6), 73–94. https://doi.org/10.1177/0263276418769733

WEXLER R. (2017), "The Odds of Justice: Code of Silence: How private companies hide flaws in the software that governments use to decide who goes to prison and who gets out", *Washington Monthly*. https://washingtonmonthly.com/magazine/junejulyaugust-2017/code-of-silence/

Des interfaces traditionnelles hommes-machines aux machines empathiques : vers une coadaptation humain-machine

Par **Laurence DEVILLERS**
Professeur en IA à Sorbonne Université/LIMSI-CNRS, membre du CNPEN et du GPAI sur le futur du travail

Introduction

Pour comprendre les particularités des robots empathiques, rappelons qu'un robot est caractérisé par trois composantes en interaction : il recueille des données grâces à ses capteurs, il les interprète grâce à ses programmes, et il peut bouger et agir sur son environnement. De plus, le robot peut avoir une apparence anthropomorphe et une capacité d'interaction langagière. On désigne par *chatbot* ou « agent conversationnel » un système de dialogue homme-machine. La robotique sociale tend à créer des robots dotés de capacités sociales, entre autres la capacité de dialoguer, qui pourraient prendre des rôles de substituts pour certaines tâches dans la société (Dumouchel et Damasio, 2016). Enfin, ces robots sociaux peuvent être dotés d'empathie artificielle.

L'empathie est une réponse émotionnelle à une situation bien particulière. C'est un trait de personnalité qui est de pouvoir ressentir une émotion appropriée en réponse à celle exprimée par une autre personne, et de bien distinguer l'émotion de l'autre (votre enfant a mal) de votre émotion (vous avez mal pour lui, mais vous ne sentez pas de souffrance physique). L'homme possède une capacité particulière à se projeter. Chez l'enfant, l'empathie affective apparaît à un an, l'empathie cognitive apparaît plus tard vers quatre ans et demi. Une empathie plus « mature », incluant le sens de la réciprocité et permettant la construction du sens moral et de la justice, est présente entre huit et douze ans.

Les robots et les *chatbots* comme Google Home ou Siri sur notre téléphone pourraient nous donner l'illusion que les machines sont empathiques, si on leur donne la capacité d'identifier les émotions de leurs interlocuteurs humains, de raisonner à partir des émotions détectées, et de générer par l'expressivité faciale, les gestes, les postures et l'expressivité acoustique des émotions. La détection des émotions d'une personne peut amener la machine à changer de stratégie de réponse. Elle peut ainsi répondre « je suis triste aussi » à quelqu'un, lorsqu'elle a détecté de la tristesse. Cette tromperie peut nous amener à croire à l'empathie des robots. Il ne s'agit pas véritablement d'empathie, car ces machines n'ont pas de « conscience phénoménale », c'est-à-dire d'expériences caractérisant le « vécu » ou le « ressenti » d'une personne.

La conscience phénoménale, contrairement à la conscience définie comme « cognition », est associée à une expérience qualitative, telle que le sentiment de plaisir ou de douleur, de chaud ou de froid, etc. La conscience phénoménale n'est donc pas réductible aux conditions physiques ou physiologiques de son apparition, et elle est indissociable de la subjectivité de l'individu. Il y a ainsi une grande différence entre les descriptions scientifiques de la conscience, qui font référence au comportement ou au fonctionnement du cerveau publiquement observables, et la conscience phénoménale propre au sujet.

Le neuroscientifique Antonio Damasio a apporté une vision nouvelle sur la manière dont les émotions se manifestent dans les interrelations étroites qu'entretiennent le corps et le cerveau

dans la perception des objets (Damasio, 1994). La notion de « corps » est centrale : l'homme ne se réduit pas à une pensée, à une conscience de soi ; il est aussi un corps, par l'intermédiaire duquel il se trouve dans le monde. Il faut lever les voiles du mystère des sciences comme l'intelligence artificielle, les neurosciences cognitives et l'intelligence affective pour mieux utiliser dans la société ces artefacts, robots ou agents conversationnels. La modélisation informatique des affects amène à se poser la question des conséquences sociétales de vivre dans un quotidien environné d'objets pseudo-affectifs (Devillers, 2017). Des principes éthiques imposant par exemple de distinguer les agents artificiels pourraient être envisagés.

L'*"affective computing"*

Le domaine de l'"*affective computing*" prend ses sources dans les travaux de Rosalind Picard (Picard, 1997), au MIT en 1997, et regroupe trois technologies : la reconnaissance des émotions des humains, le raisonnement et la prise de décision en utilisant les informations recueillies, et la génération d'expressions émotionnelles. Ce domaine est par essence pluridisciplinaire. La reconnaissance des messages sociaux véhiculés par les visages et les voix, et en particulier les expressions émotionnelles, est un élément indispensable à la communication avec les humains et à l'insertion dans toutes les sociétés.

De façon consensuelle, l'émotion est définie comme une réaction à un événement, une situation, réel(le) ou imaginaire, comprenant plusieurs facettes ou composantes. Trois composantes sont généralement acceptées comme constituantes essentielles de la réaction émotionnelle. Il s'agit du sentiment subjectif (vécu émotionnel), de la réaction physiologique et de l'expression émotionnelle (faciale, vocale ou posturale). Dans un contexte de communication, les expressions sont transformées en fonction d'un ensemble de règles socio-culturelles. Ces règles varient d'une culture ou d'un groupe social à l'autre, dans des contextes « objectivement » similaires.

Des situations sociales peuvent exiger la suppression de certaines expressions alors que d'autres situations au contraire exigent de montrer, voire d'exagérer des expressions spécifiques. Masquer une expression spontanée qui ne serait pas désirable dans un contexte social donné est également possible. Le contexte social peut influencer les expressions émotionnelles en fonction de la position et des objectifs de l'émetteur dans la situation. Dans une situation de communication, un individu peut utiliser ses expressions émotionnelles de manière à influencer – plus ou moins (in)consciemment et plus ou moins (in)volontairement – les réactions de ses interlocuteurs.

Les machines vont de plus en plus interagir vocalement avec nous dans la vie de tous les jours. Les agents conversationnels et les robots sociaux peuvent déjà embarquer des systèmes de détection, de raisonnement et de génération d'expressions affectives qui, même avec des erreurs importantes, peuvent interagir avec nous. Ils envahissent maintenant nos sphères privées. On dénombre aux USA jusqu'à six enceintes Alexa ou Google Home par foyer, une par pièce. Le marché est énorme, ces machines pourraient nous accompagner au quotidien, pour surveiller notre santé, nous éduquer, nous aider et nous amuser, bref pour s'occuper de nous. Pour ces tâches, la machine est créée en vue d'être une sorte de compagnon numérique, assistant ou surveillant.

La communication avec les machines est avant tout un échange d'information avec trop souvent de l'« incommunication ». L'incommunication peut se produire même dans des circonstances où de l'information est communiquée, si l'information ne contient pas de message ou si le récepteur, par exemple le robot, ne peut pas décoder l'information contenue dans le message. Les machines sont loin d'avoir des capacités sémantiques suffisantes pour converser et partager des idées, mais elles pourront bientôt détecter notre malaise, notre stress, peut-être certains de nos mensonges.

L'empathie des humains pour les machines

La "*media equation*" de Reeves et Nass (1996) explique que nous appliquons les mêmes attentes sociales lorsque nous communiquons avec des entités artificielles, et que nous assignons inconsciemment à celles-ci des règles d'interaction sociale.

L'anthropomorphisme est l'attribution des caractéristiques comportementales ou morphologiques de vie humaine à des objets. Avec ce réflexe, à la fois inné et socialement renforcé, un objet qui semble être dans la douleur peut inspirer de l'empathie. Des études expérimentales ont montré la projection de réactions affectives et probablement empathiques envers des entités artificielles, par exemple des robots jouets volontairement endommagés.

Les chercheurs ont constaté que les humains ressentaient de l'empathie envers des robots maltraités, certes de moindre intensité qu'envers des humains maltraités, mais cette empathie n'existe pas envers des objets inanimés. Les recherches récentes, grâce à l'imagerie cérébrale, indiquent que les individus répondent de façon étonnamment semblable aux images émotionnelles des humains et à celles des entités artificielles. Si nous représentons les entités artificielles comme des humains, alors il n'est peut-être pas surprenant que nous réagissions avec émotion envers les agents artificiels comme nous réagissons envers les humains, cependant il n'est pas clair que des représentations de même complexité soient attribuées aux robots.

Les robots Nao, Pepper et Romeo sont des robots sociaux qui peuvent être aussi des miroirs de nos émotions, mais ils ne sont pas à proprement parler empathiques. Fan Hui, champion de Go qui a entraîné AlphaGo pour Google DeepMind, expliquait que jouer contre une machine, c'était un peu jouer contre soi-même, car on projetait ses émotions sur la machine qui les renvoyait comme un miroir. Le robot PARO le phoque, développé dès 1993 au Japon, a été commercialisé au Japon en 2005, puis aux États-Unis en 2009 (certification FDA en tant que robot thérapeutique).

Quand PARO est utilisé en EHPAD auprès de personnes âgées, ses réactions ne sont pas empathiques, elles ressemblent plus à des comportements expressifs d'animal de compagnie. Lors de nos tests en EHPAD (Garcia *et al.,* 2017) avec les robots Nao et Pepper qui détectaient les émotions et s'adaptaient en conséquence, les personnes âgées que nous avons rencontrées (une bonne cinquantaine de personnes de moyenne d'âge de 85 ans), qui n'étaient pas sous tutelle, ont eu des réactions de curiosité et d'amusement. Elles n'ont cependant pas considéré que le robot les « comprenait », mais qu'il pouvait détecter certains de leurs comportements.

Un enfant de deux ans sait que son doudou n'est pas vivant, et pourtant il lui parle comme si celui-ci l'était. Les psychologues parlent des poupées et doudous comme d'« objets transitionnels ». L'objet transitionnel est donc un objet privilégié, choisi par l'enfant, généralement doux au toucher. Il permet au bébé de lutter contre l'angoisse, il est la première possession. Il n'est perçu ni comme faisant partie de la mère, ni comme étant un objet intérieur. Il permet le cheminement de l'enfant du subjectif vers l'objectif. PARO le robot en peluche peut être vu comme un objet transitionnel. Les robots sociaux sont également avant tout des objets techniques qui enregistrent nos données pour les transférer par exemple à un médecin.

Les robots assistants de vie s'insèrent aussi dans un écosystème qui comprend de nombreux acteurs : la famille, les aides-soignants, les médecins.

Les troubles de comportements face aux robots sont étudiés en psychologie et en psychiatrie, notamment par Serge Tisseron, psychiatre (Tisseron, 2015). Kate Darling, chercheuse au Media lab du MIT, étudie les réactions empathiques des personnes devant des robots, notamment en cas de maltraitance des robots, avec l'idée qu'il faudrait accorder aux robots une protection juridique comme on l'a fait pour les animaux (Darling, 2016). Pourtant, les animaux sont des organismes

vivants qui peuvent ressentir des émotions et souffrir, ce qui n'est absolument pas le cas des robots qui ne font que simuler des émotions.

Les émotions des machines

Donner aux machines des capacités d'interprétation et de simulation émotionnelle est indispensable pour construire des systèmes capables d'interagir socialement et de mieux communiquer avec les humains. Les applications dans le but d'aider les personnes dépendantes, dans le grand âge ou pour différentes pathologies dégénérescentes sont nombreuses. La sphère des émotions que l'on pouvait penser propre à l'humain envahit les machines, qui se rapprochent des capacités humaines. Dans certains cas, une hésitation, un souffle de la machine donne l'impression qu'elle est en « vie ». Donal Davinson, philosophe américain, décrit le monisme anomique comme « l'union exprimée par deux langages différents sans traduction ». Il n'y a pas de raison causale entre l'esprit et le corps. Plus la machine a l'air fragile, plus on peut l'humaniser et être ému devant elle, même si à proprement parler elle n'est pas empathique.

Le philosophe Spinoza et plus particulièrement son ouvrage *L'Éthique* (1677) est une source d'inspiration pour expliquer le monde d'aujourd'hui et les relations entre le corps et l'esprit. Spinoza explique que l'organisme se fabrique lui-même. Les affects et les haines, lieux par excellence où sont unis le corps et l'esprit, sont sources d'aliénation si nous les subissons et/ou de liberté si nous en comprenons les mécanismes sous-jacents. Grâce à l'exploration du cerveau, nous pouvons aujourd'hui prouver les déclarations du philosophe qui étaient contraires au sens commun à son époque. Ainsi, les expressions corporelles précèdent le sentiment. Corps et esprit sont mélangés, nous apprend Spinoza. Mais les machines n'ont pour l'instant pas de corps au sens de viscères, d'hormones, de peau ; elles n'ont pas d'intention, de plaisir et de désir propres, c'est-à-dire pas de « *conatus* » au sens entendu par Spinoza. Le vivant peut par conséquent être défini comme autonome et ayant la possibilité de se reproduire.

À l'heure actuelle, la relative autonomie des robots est toujours programmée par l'humain. La faculté d'apprentissage programmée peut offrir plus ou moins de liberté à la machine. Donner à un robot la capacité d'apprendre seul, en interaction avec l'environnement et les humains, est le Graal des chercheurs en intelligence artificielle. Si les robots apprennent seuls, il sera souhaitable de leur enseigner les valeurs communes et morales de la vie en société. Cette faculté constitue cependant une rupture technologique et juridique, et soulève de nombreuses questions éthiques. Ces robots peuvent être, d'une certaine manière, créatifs et autonomes dans leurs prises de décision, si on les programme pour cela.

Vouloir recopier l'intelligence de l'homme sur une machine est très narcissique, car que connait-on de notre intelligence ? Nous ne connaissons pas le substrat de la pensée et n'avons pas conscience que certains de nos organes sont autonomes, nous ne sommes conscients que d'une petite partie de nos perceptions et de notre activité cérébrale. Il n'existe pas de terme plus polysémique et sujet à interprétation que celui de « conscience » : il évoque pour certains la conscience de soi, pour d'autres la conscience du prochain, ou encore la conscience phénoménale, la conscience morale, etc.

Avec une conception philosophique matérialiste de la vie, on peut considérer que l'ordinateur et le cerveau humain sont des systèmes comparables, capables de manipuler des informations. Les modélisations numériques les plus performantes, comme le *deep learning* (apprentissage profond), s'appuient sur une modélisation simplifiée du neurone (neurone formel), intégrée dans une machine à états discrets simulée sur ordinateur. Le nombre de couches cachées de l'architecture du modèle correspond à la profondeur. Pour l'instant, nous sommes très loin de la complexité du vivant !

Les systèmes actuels d'intelligence artificielle ont la capacité de calculer des corrélations de faits, avec les approches d'apprentissage profond par exemple, de prendre des décisions et d'apprendre, mais sans en avoir conscience. Certains prototypes de robots ont pourtant déjà des embryons de niveau de « conscience » comparables à ceux que décrit Stanislas Dehaene. Ils sont simulés par des mécanismes de partage de connaissances et d'introspection. Pour autant, ces machines ne sont pas conscientes comme peut l'être un humain, elles n'ont ni conscience morale ni conscience phénoménale associée à une expérience qualitative telle que la sensation de chaud ou de froid, le sentiment d'anxiété, ..., car elles n'ont pas de viscères ni de ressenti, à moins, là encore, de les simuler.

Les premiers robots apprenants et communicants, dotés de capteurs de douleur et de plaisir (Asada, 2015), interagissent par des procédés simples : des capteurs sur leur corps, une caméra et un microphone qui leur permettent d'associer un visage et une voix aux signes expressifs qu'ils reçoivent. Leurs indicateurs d'expression (sons émis, diodes) permettent aux humains de comprendre leurs états. Pour produire des effets rapides, il faut des actions physiques : une caresse sur la tête, et ils associent un critère positif à la personne qu'ils voient ; une tape sur la tête, et ils lui associent un critère négatif. Ces machines apprennent et adaptent leur comportement aux contextes dans lesquels ils se trouvent. Mais les interprétations sont encore très limitées. Par exemple, un robot assistant d'étudiants en chirurgie dentaire devrait informer au mieux sur les grimaces de douleur, et anticiper des gestes afin d'alerter sur la proximité d'un nerf et une possible souffrance. Faut-il que les robots s'approchent le plus possible des humains ? Une conscience artificielle, dotée de sentiments, de pensées et de libre arbitre sans programmation humaine a encore peu de chance d'émerger spontanément avec les architectures actuelles d'ordinateurs.

Conclusion

La vie au quotidien avec des robots pourrait entraîner des risques sociaux à long terme qu'il faut anticiper pour tirer bénéfice de ces machines. L'addiction et l'isolement, ainsi que le report d'autonomie sur la machine, la confusion entre la machine et l'humain sont des déviances dont il faut se préoccuper. Un des risques, particulièrement pour les personnes fragiles, est d'oublier qu'un robot est connecté et programmé. La capacité d'un robot de s'adapter à son propriétaire humain pourrait bien être utilisé pour lui faire accomplir certains choix plutôt que d'autres, notamment pour l'aider à mieux gérer les déviances de comportements de ce dernier (pathologie sexuelle, addiction à la drogue), mais aussi pour des causes moins louables dans le domaine de la consommation. Un autre risque est d'oublier qu'un robot ne ressent rien, n'a pas d'émotions, n'a pas de conscience et n'est pas vivant. Il est possible de ressentir de l'empathie pour un robot et de parler de souffrance pour un robot. Il est important que les personnes âgées, qui peuvent mettre leur vie en danger pour venir en aide à leur robot, se rendent compte qu'un robot ne souffre pas même s'il tombe, il faut qu'elles soient conscientes que ce n'est qu'un objet programmé.

Les robots empathiques soulèvent de nombreuses questions éthiques, juridiques et sociales. Ces questions prégnantes ne sont évoquées que depuis peu. Les progrès spectaculaires du numérique permettront un jour d'améliorer le bien-être des personnes, à condition de réfléchir non à ce que nous pouvons en faire, mais à ce que nous souhaitons en faire. Un certain nombre de valeurs éthiques sont importantes : la déontologie et responsabilité des concepteurs, l'émancipation des utilisateurs, l'évaluation, la transparence, l'explicabilité, la loyauté et l'équité des systèmes, enfin l'étude sur le long terme de la « coadaptation » humain-machine (la machine s'adaptera à l'humain et l'humain à la machine).

Le contrôle par des humains sera toujours primordial. Il est nécessaire de développer des cadres éthiques pour les robots sociaux, notamment dans le domaine de la santé, et de comprendre le

niveau de complémentarité humain-machine. Nous avons besoin de démystifier, de former à l'intelligence artificielle et de remettre au centre de la conception de ces systèmes robotiques les valeurs de l'humain.

Bibliographie

ASSADA M. (2019), "Artificial Pain: empathy, morality, and ethics as a developmental process of consciousness", *Philosophies* 2019, 4, 38; doi: 10.3390.

BOSTROM N. (2014), *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press.

DAMASIO D. (1994), *L'Erreur de Descartes*, Éditeur Odile Jacob.

DARLING K. (2016), « Faut-il accorder une protection juridique aux robots de compagnie ? », dans le livre d'Alain Bensoussan, Yannis Constantinides, Kate Darling, Jean-Gabriel Ganascia et Olivier Tesquet, *En compagnie des robots,* Éditeur Premier parallèle.

DEHAENE S. (2014), *Le code de la conscience,* Éditeur Odile Jacob.

DEVILLERS L. (2017), *Des robots et des hommes : mythes, fantasmes et réalité,* Éditeur Plon.

DUMOUCHEL P. & DAMIANO L. (2016), *Vivre avec les robots*, Éditeur Seuil.

GARCIA M., BECHADE L., DUBUISSON DUPLESSIS G., PITTARO G. & DEVILLERS L. (2017), "Towards Metrics of Evaluation of Pepper Robot as a Social Companion for Elderly People", International Workshop on Spoken Dialogue Systems (IWSDS), 8 p.

PICARD R. (1997), *Affective computing*, MIT Press.

SPINOZA B. (1677), *L'Éthique,* Livre III, Proposition II, et SCOLIE, (Point-Essais), trad. Bernard Pautrat, pp. 207-209.

REEVES B. & NASS C. (1996), *The Media Equation,* CSLI Publications, Stanford University.

TISSERON S. (2015), *Le jour où mon robot m'aimera : vers l'empathie artificielle,* Éditeur Albin Michel.

Résumés

peu la possibilité d'une telle éventualité. S'il est donc peu probable qu'une IA quelconque pilote un jour les décisions stratégiques d'une entreprise, son utilisation dans les stratégies d'entreprise en revanche est déjà une réalité qui modifie l'architecture des ressources et des compétences au sein de l'entreprise. Cette nouvelle architecture de la création de valeur nécessite donc des réorganisations internes pour la déployer au sein des stratégies métiers. Par la nature des décisions que l'IA automatise, il devient impératif pour les entreprises de se doter d'un organe de gouvernance définissant la doctrine d'usage de telles technologies.

To analyze the concept of interpretability associated with an ML algorithm, a distinction is made between "how" (How does a black box or a very complex algorithm work?) and "why" (Why does an algorithm produce such-and-such a result?). These questions appeal to many actors: users, professionals, regulators, etc. Using a formal, standardized framework, existing solutions are indicated by specifying which elements in the supply chain are impacted when answers are provided to the previous questions. This presentation, by standardizing notations, allows for a comparison of different approaches so as to highlight the specificity of each (their objectives and processes). This study is not exhaustive — the subject is far from closed…

L'arrivée de l'ensemble de technologies qualifiées parfois « d'informatique cognitive », d'IA, de *big data*… (Shivam *et al.*, 2018) pourrait bouleverser l'approche de la performance des entreprises. Plus qu'une nouvelle manière d'analyser la performance (à travers l'apparition de nouveaux indicateurs), c'est un nouveau rapport aux données chiffrées que ces technologies induisent. Mais ces approches ne sont pas sans risque. Afin d'éviter les dangers de modèles algorithmiques qui seraient des boites noires et afin de répondre aux enjeux de schémas d'interprétation mis sous tension, c'est une transformation des métiers (et principalement de celui des contrôleurs de gestion) et des organisations qui est attendue. Les chiffres étant des conventions, des construits sociaux, toute transformation des modalités de production des chiffres implique des transformations des systèmes sociaux sur lesquels ils agissent.

Aujourd'hui omniprésente, l'intelligence artificielle (IA) a un impact, par sa transversalité, sur tous les secteurs, à tel point que rares seront les activités humaines dont elle restera exclue. Levier de croissance de nature à modifier en profondeur les modes de production et les modèles économiques existants, elle préfigure, aux yeux de certains, de nouveaux types de rapports sociaux qui ne seraient pas purement humains. D'où l'importance de concevoir une stratégie de régulation de l'IA par l'éthique, mais aussi par le droit. Et de mettre en place un véritable écosystème juridique propre aux algorithmes. D'autant que l'implication des plus grands acteurs de l'économie numérique mondiale, l'importance des enjeux financiers, l'engouement de la recherche et les questions d'acceptabilité sociale constituent une assise particulièrement solide à l'avènement d'un véritable droit de l'intelligence artificielle.

73 Algorithmes et droit pénal : quel avenir ?

Elise BERLINSKI, Imane BELLO et Arthur GAUDRON

Les algorithmes d'intelligence artificielle (IA) et plus spécifiquement de *machine learning* (ML) se déploient dans le domaine du droit pénal. Ils permettent la création de sujets numériques comme prismes de perception d'un sujet référent réel. Alors que le lien entre le sujet référent et le sujet numérique n'est pas évident, les informations que ce dernier offre peuvent orienter une enquête ou influencer l'appréciation d'un juge. Il semble donc urgent de comprendre plus précisément le prisme par lequel ces méthodes font percevoir la personne physique au centre de l'enquête ou du jugement. Pour cela, nous décrivons les étapes de construction du sujet numérique, des données utilisées aux algorithmes appliqués, dont les spécificités propres se présentent comme autant de réflexions kaléidoscopiques. Nous considérons enfin les degrés de transformation que l'utilisation d'un sujet numérique créé par des algorithmes de ML pourrait impliquer pour le droit pénal.

78 Des interfaces traditionnelles hommes-machines aux machines empathiques : vers une coadaptation humain-machine

Laurence DEVILLERS

Le domaine de l'"*affective computing*" est par essence pluridisciplinaire, en regroupant trois technologies : la reconnaissance des émotions des humains, le raisonnement et la prise de décision en utilisant les informations recueillies, et la génération d'expressions émotionnelles. La sphère des émotions que l'on pouvait penser propre à l'humain envahit les machines qui se rapprochent des capacités humaines. Dans le cas de maltraitance d'un robot notamment, l'être humain peut ressentir de l'empathie envers un agent artificiel, alors que ce dernier ne fait que simuler des émotions et ne peut pas souffrir. La machine est programmée, sans intention ni désir propres, mais il faut le rappeler. L'arrivée des robots empathiques soulève depuis peu de nombreuses questions éthiques, juridiques et sociales.

Abstracts

The transportation sector has been making innovations centered on its industrial aspects for decades now. After several successive waves of automation, it is now up to artificial intelligence to bring to light new potentials for improving the quality of services and the user experience, which will be more customized and inviting. These new capacities, which are opening for the RATP group (the Paris Area Transit Authority), are vectors of change and innovation for building models of new trends and providing knowledge that should be made concrete and operational. The subtle match between expertise in business processes, mathematic, computer science and data is the key to success.

Thanks to advances in design and safety, industrial workstations now offer the possibility for robots and operators to share the same workspace and interact with each other. It thus becomes necessary to rethink the organization of workstations so that operatives can collaborate with "cobots" (collaborative robots). Examples are provided to illustrate how artificial intelligence techniques can facilitate the design of these new workstations and improve human-robot interactions.

The complex relation between people and driving extends beyond the mere use of a technological object. The introduction of artificial intelligence (AI) opens unprecedented perspectives and raises, again, questions about the understanding and acceptability of a new technology. The gradual rollout of driverless vehicles is described along with the current state-of-the-art techniques being implemented; the prospects for this trend are analyzed; and the issues to be addressed, pointed out. The challenges are, of course, technological; but it will also be necessary to reinvent a new mobility. How will the time be used that is now perceived to be wasted on the uninteresting task of driving in current traffic conditions?

Given the rapid development over the past decade of methods qualified as "artificial intelligence" (AI), questions arise about how these methods might fit into a firm's strategies, or even replace them. This view overlooks the aspects of strategy-making that are marked with a high degree of uncertainty and many an ambiguity. The limitation inherent in building tools for decision-making that massively rely on sets of data restricts somewhat the possibility of this happening. Although it is unlikely that AI will some day steer a firm's strategic decisions, its use in corporate strategies is already a reality that is modifying the

create kaleidoscopic reflections. What changes does the creation of the digitized subject via ML algorithms imply for penal law?

78 From traditional man-machine interfaces to empathic machines: Toward a coadaptation

Laurence DEVILLERS

The multidisciplinary field of "affective computing" groups three sorts of technology: the recognition of human emotions; deliberation (reasoning and the making of decisions based on the information gathered); and the generation of emotional expressions. The sphere of emotions, which we might imagine as being specifically human, is encompassing machines as the latter are increasingly endowed with human capacities. When a robot is mistreated, a human being might feel empathy toward this artificial agent, which is merely simulating emotions and cannot suffer. The machine is programmed; it lacks intentionality and has no desires of its own. The coming of empathic robots raises ethical, legal and social questions…

Ont contribué à ce numéro

Yohan AMSTERDAMER est responsable du « programme intelligence artificielle » au sein du groupe RATP. Ingénieur diplômé des Mines de Paris, détenteur d'un master à l'ESSEC, il a débuté sa carrière dans l'ingénierie des transports, où il a évolué en tant que responsable performance financière et opérationnelle, un poste transverse qui lui a permis d'acquérir une solide expérience en matière de gestion de projets et de conduite du changement. Il rejoint la RATP en août 2018 pour structurer et activer les leviers de l'IA au service de l'excellence opérationnelle et l'expérience client.
→ ***L'intelligence artificielle en milieu industriel, levier de transformation et facteur d'innovation du groupe RATP***

Julien BARNU est un ancien élève de l'École polytechnique et ingénieur en chef des Mines. Il rejoint en 2013 l'agence nationale de la sécurité des systèmes d'information (ANSSI) en tant que chef de projet, en charge de la politique de cybersécurité des activités d'importance vitale, puis en tant que directeur de cabinet du directeur général, Guillaume Poupard, de janvier 2016 à juillet 2018. Il assure depuis le 1er août 2018 la fonction de conseiller industrie et numérique du secrétaire général de la défense et de la sécurité nationale (SGDSN).
→ ***Intelligence artificielle et sécurité nationale***

Imane BELLO est avocate à la Cour. Elle intervient principalement en matière de droit pénal du numérique, protection des données à caractère personnel, compliance et intelligence artificielle (gouvernance, gestion des risques et éthique). Elle est également chargée d'enseignement en éthique et politique des systèmes d'intelligence artificielle à l'Institut d'études politiques de Paris (Sciences Po Paris). Elle intervient régulièrement lors de conférences internationales (Unesco, OCDE) sur les sujets liés à la gouvernance politique et juridique des systèmes d'intelligence artificielle, au droit numérique et au respect des droits humains. Imane Bello est titulaire d'une maîtrise en gouvernance mondiale de l'École de droit de Sciences Po Paris, ainsi que de licences en sciences politiques (Université Libre de Berlin), en droit (Université Nancy II) et en sciences sociales (Sciences Po Paris).
→ ***Algorithmes et droit pénal : quel avenir ?***

Salima BENHAMOU (PhD) est économiste à France Stratégie. Ses domaines d'expertise portent sur l'avenir du travail, l'intelligence artificielle, les enjeux liés aux transformations des organisations sur les compétences, le management et l'innovation, ainsi que sur la gouvernance des entreprises, la participation des salariés et la RSE. Elle a notamment publié un rapport prospectif sur l'avenir du travail en 2030 et une étude sur les enjeux et les défis pour la France des organisations apprenantes, en matière de qualité du travail et de l'emploi et de diffusion des innovations. Elle a également coordonné le rapport « Intelligence artificielle et travail » remis à la ministre du Travail et au ministre d'État du Secteur numérique, et a été membre du groupe d'experts internationaux du G7 en 2018, sous la présidence canadienne, pour alimenter les discussions des ministres de l'Innovation et du Travail sur le thème principal « Avenir du travail et IA : des compétences pour l'économie moderne ».
→ ***Intelligence artificielle et travail : le défi organisationnel***

Alain BENSOUSSAN est avocat à la Cour d'appel de Paris et précurseur du droit des technologies avancées. Il a fait de l'élaboration de concepts nouveaux l'une de ses marques de fabrique : domicile virtuel, droits de l'homme numérique, vie privée résiduelle, etc. En 2012, après

avoir créé Lexing®, premier réseau international qui fédère des avocats en droit du numérique et des technologies avancées, il lance au sein de son cabinet un département sur le droit des technologies robotiques, y voyant « la reconnaissance par le droit d'une mutation technologique au moins aussi importante que l'ont été l'informatique et les réseaux sociaux au XXe siècle ». Aux yeux de cet infatigable explorateur de nouveaux domaines du monde numérique, il était temps de créer un droit des robots, les dotant d'une personnalité et d'une identité juridique pour en faire, demain, des sujets de droit : « Avec l'introduction d'une intelligence artificielle, les robots ne sont pas de simples automates. Ils ont des capacités grandissantes qui les amènent à collaborer avec les hommes. » Alain Bensoussan est également président et fondateur de l'association du droit des robots (ADDR).
→ ***Quelle régulation juridique pour l'intelligence artificielle ?***

Côme BERBAIN est directeur de l'Innovation du groupe RATP et directeur du « programme véhicule autonome ». Ingénieur du corps des Mines, docteur en cryptographie, il a alterné entre entités privées (Orange, Trusted Logic) et publiques (ministère de la Défense, ANSSI, direction interministérielle du numérique - DINUM, dans les domaines de la transformation numérique et de la cybersécurité. En 2017 et 2018, il est conseiller au cabinet du secrétaire d'État chargé du Numérique, où il porte les sujets de transformation numérique de l'État et de confiance numérique, avant de devenir Chief Technology Officer de l'État à la direction interministérielle du numérique en 2019. Il rejoint la RATP en novembre 2019.
→ ***L'intelligence artificielle en milieu industriel, levier de transformation et facteur d'innovation du groupe RATP***

Elise BERLINSKI finit actuellement sa thèse en sciences de gestion à l'ESCP. Diplômée de l'ENSIMAG et de l'Imperial College London en 2013, elle a commencé à travailler sur les algorithmes d'apprentissage, plus spécifiquement de théorie des jeux stochastiques, lors d'un stage de recherche dans le département Electrical and Electonic Engineering. À la suite de cela, elle a pratiqué pendant quatre ans comme Data Scientist dans l'industrie, travaillant dans le domaine des paris sportifs, puis de la cybersécurité, tout en menant diverses missions de conseil. À présent, elle s'intéresse à la manière dont des imaginaires sociaux participent à modeler les technologies et leurs usages. Par ailleurs, elle porte un intérêt spécifique aux processus de construction des nombres, en particulier numériques, et à la manière dont ceux-ci sont utilisés.
→ ***Algorithmes et droit pénal : quel avenir ?***

Nicolas BERLAND est professeur à l'Université Paris Dauphine où il dirige l'Executive MBA. Il a été directeur de DRM (Dauphine Recherches en Management) jusqu'en décembre 2018. Il est spécialisé en contrôle de gestion et dans le pilotage des organisations. Son approche est résolument orientée vers l'étude de l'articulation du contrôle de gestion aux processus stratégiques et aux processus de management de l'entreprise. Ses recherches l'ont conduit à étudier l'histoire des pratiques de contrôle de gestion, la financiarisation du contrôle, les expériences de « gestion sans budget », les pratiques de contrôle en RSE et l'implantation de contrôle de gestion en milieu public (la Police nationale par exemple). Il a notamment publié *Contrôle de gestion, perspectives stratégiques et managériales* chez Pearson et *Le contrôle de gestion* aux Éditions « Que sais-je ? ». Il est membre du conseil scientifique de la DFCG (association des directeurs financiers et contrôleurs de gestion).
→ ***Intelligence artificielle et contrôle de gestion : un rapport aux chiffres revisité et des enjeux organisationnels***

Yves CASEAU est le directeur des systèmes d'information du Groupe Michelin depuis octobre 2017, après avoir été le directeur digital du Groupe AXA. Il a été le directeur général adjoint Technologies, Services et Innovation de Bouygues Telecom de 2007 à 2013. Il a consacré la première partie de sa carrière scientifique au génie logiciel, à la programmation par objet et à l'intelligence artificielle, puis il s'est tourné vers la recherche opérationnelle dans les années 90, à Telcordia (USA) puis dans le Groupe Bouygues. Ancien élève de l'ENS (Ulm), Yves Caseau est titulaire d'un doctorat et d'une habilitation à diriger les recherches en informatique, ainsi que d'un MBA du Collège des Ingénieurs. Il est membre de l'Académie des Technologies et auteur de quatre livres chez Dunod : *Urbanisation, BPM et SOA* (2005), *Performance du Système d'Information* (2007), *Processus et Entreprise 2.0* (2011) et *L'approche lean de la transformation digitale* (2020).
→ ***Le futur du travail en présence de formes artificielles d'intelligence***

Arnaud de LA FORTELLE est professeur à Mines ParisTech, où il dirige le Centre de Robotique. Après des études à l'École polytechnique et aux Ponts-et-Chaussées (incluant un doctorat sur les systèmes stochastiques), il travaille à l'Inria avant de rejoindre Mines ParisTech en 2006. Il est *visiting professor* en 2017-2018 à UC Berkeley, où il enseigne le contrôle des systèmes distribués et travaille avec l'équipe Berkeley Deep Drive. Il est spécialiste des systèmes coopératifs (communication, perception, contrôle et certification mathématique) appliqués au véhicule autonome comme aux systèmes de transports complets. Arnaud de La Fortelle a présidé la commission scientifique ANR mobilité et systèmes urbains durables et est expert auprès de la Commission européenne. Il est titulaire des chaires Drive for All sur le véhicule autonome (PSA, Valeo et Safran, avec EPFL, Shanghai Jiao Tong et Berkeley) et logistique urbaine (Renault, La Poste, Pomona, ville de Paris et Ademe).
→ ***Introduction***
→ ***Une IA ou des IA ? Représentations et relations avec les IA***

Guillaume DEVAUCHELLE est vice-président Innovation et Développement scientifique du groupe Valeo depuis 2012. Il était précédemment directeur R&D du Groupe depuis 2004. Il a rejoint le Groupe Valeo en 2000 suite au rachat des activités de Sylea, dont il était DGA. Il a fait toute sa carrière dans l'équipement Aéronautique et Automobile. Il est diplômé de l'Ecole centrale de Paris (1981). Il est membre de l'Académie des Technologies. Il s'est particulièrement investi dans le rapprochement de la recherche publique et privée en tant que :

- Actuellement :
  - vice-président de l'ESIGELEC (école supérieure d'ingénieurs),
  - président fondateur, membre du Bureau de VEDECOM,
  - administrateur de l'ANRT (association nationale de la recherche et de la technologie), président par intérim,
  - membre du CTA (comité technique automobile) et du CRA (conseil de la recherche automobile),
- Après avoir été :
  - administrateur de l'UVSQ (Université de Versailles Saint-Quentin),
  - membre du Conseil de perfectionnement de IFPEN School,
  - administrateur du CETIM (centre d'étude et technique des industries mécaniques)
  - vice-président de la SIA (société des ingénieurs de l'automobile),
  - vice-président du Pôle de Compétitivité de Moveo,
  - administrateur de l'UTAC (union technique et du cycle),
- Également administrateur de *start-up* :
  - Aledia (Nano LED), Navya (véhicule autonome).

→ ***La conduite automatisée, intelligences artificielles et humaines, quelles interactions ?***

Laurence DEVILLERS, professeur en IA à Sorbonne-Université (depuis 2011), dirige l'équipe « Dimensions affectives et sociales des interactions parlées » (CNRS-LIMSI). Elle est responsable de la chaire HUMAAINE : HUman-MAchine Affective Interaction & Ethics (Saclay) sur l'informatique émotionnelle, le *nudge* des agents/robots conversationnels et l'éthique. Elle est membre du comité national pilote éthique du numérique (CCNE-numérique) et du Global Partnership on AI (juin 2020) sur le futur du travail. Elle a encadré 11 thèses (+ 4 en cours), est l'auteur de plus de 170 publications (h-index : 38) et d'essais : *Des robots et des hommes* (PLON, 2017), *Les robots « émotionnels »* (Éditions de l'Observatoire, 2020) et *La souveraineté numérique dans l'après-crise* (Éditions de l'Observatoire, 2020). Elle collabore avec le MILA (Canada) et est impliquée dans l'institut Human-Centered Artificial Intelligence (https://www.humane-ai.eu/).
→ ***Des interfaces traditionnelles hommes-machines aux machines empathiques : vers une coadaptation humain-machine***

Arthur GAUDRON est titulaire d'un doctorat en mathématiques et systèmes de Mines ParisTech et de l'Université Paris Sciences et Lettres. Chercheur au Centre de Robotique, il s'intéresse aux techniques de modélisation de systèmes complexes, mais aussi à l'intégration des parties prenantes dans le processus de modélisation pour l'analyse d'un système ou pour la prise de décision. Il s'agit d'explorer les possibilités offertes par les algorithmes d'apprentissage automatique (*machine learning*) pour modéliser des systèmes complexes avec une précision accrue. En général, cependant, l'utilisation de ces techniques se fait au détriment de la compréhension du modèle, qui est pourtant nécessaire au débat et à la confiance. Ainsi, l'intégration des parties prenantes dans le processus de modélisation vise à capturer leur savoir et à le retranscrire dans le modèle, grâce à leurs interactions avec celui-ci, mais aussi dans le même temps de leur permettre de s'en saisir et d'élargir leur connaissance du système.
→ ***Algorithmes et droit pénal : quel avenir ?***

Dominique GUÉGAN is currently emeritus professor of mathematics at the Université Paris 1 Panthéon-Sorbonne and associate professor at University Ca'Foscari in Venezia. Her domains of research are: financial regulation, fintech technology (Blockchain, big data, HFT), non-linear econometrics modelling, extreme value theory and risk measures in finance, pricing theory in incomplete markets, and deterministic dynamical systems. She has published 11 books in statistics theory, time series and finance, participated for chapters in 30 books, published more than 130 academic papers, and supervised 30 PhD students. She is involved in several European projects and collaborations with French, European, Chinese, and North American universities on innovation fields, including Blockchain and Artificial intelligence.
→ ***A Note on the Interpretability of Machine Learning Algorithms***

Henri ISAAC, docteur en sciences de gestion, est maître de conférences à PSL, Université Paris-Dauphine, et chercheur au sein de Dauphine Recherches en Management (CNRS, UMR 7088). Il a été directeur de la Recherche et directeur académique à Neoma Business School (2009-2012), et vice-président « Transformation numérique » de l'Université Paris-Dauphine (2014-2016). Il dirige le parcours management télécoms et médias du Master « Systèmes, Information, Réseaux, Numérique ». Il est président du Think Tank Renaissance Numérique.

Auteur de plusieurs ouvrages, *Modèles d'affaires des plateformes* (Vuibert, 2021, à paraître), *E-commerce. Vers le commerce connecté* (4ème édition, 2017, Pearson France), et co-auteur de *Marketing digital* (7ème édition, 2020, Pearson), il a également publié de nombreux articles dans des revues académiques comme *Journal of Business Strategy, European Journal of Information Systems*, *International Journal of Innovation and*

Technology Management, *International Journal of Mobile Communications*, *Revue Française du Marketing*, *Système d'Information & Management*, *Revue Française de Gestion*.
→ ***Stratégie et intelligence artificielle***

Antoine LAFAY est ingénieur, diplômé de Supélec et d'un MBA au Collège des Ingénieurs. Il a commencé sa carrière chez Alstom Transport, en R&D au sein de la branche signalisation ferroviaire, où il a développé la nouvelle génération de métro sans conducteur. Il a ensuite intégré Valeo en 2015, pour diriger les développements série des ADAS pour les constructeurs automobiles français. Depuis 2018, il dirige la Recherche et l'Innovation sur le véhicule autonome pour Valeo.
→ ***La conduite automatisée, intelligences artificielles et humaines, quelles interactions ?***

Christian MOINARD est professeur à Audencia, où il est responsable de la majeure Contrôle de gestion/Audit. Il enseigne à Audencia et dans plusieurs institutions internationales (CFVG Viet Nam, ESAA Algérie…) depuis plus de 10 ans, après plusieurs années de pratiques dans les domaines de l'audit, du conseil et du contrôle de gestion. Ses enseignements et ses travaux l'ont amené à étudier plus spécifiquement les relations entre le contrôle de gestion et la stratégie, le contrôle de gestion public, mais aussi la place de l'informatique cognitive sur le rôle des contrôleurs de gestion et les enjeux cognitifs de cette fonction.
→ ***Intelligence artificielle et contrôle de gestion : un rapport aux chiffres revisité et des enjeux organisationnels***

Vincent WEISTROFFER a obtenu son diplôme d'ingénieur civil en 2011 (P08) et son diplôme de docteur en 2014 (P11), délivrés par l'École nationale supérieure des Mines de Paris. Au cours de sa thèse, dans le cadre de la chaire Robotique & Réalité Virtuelle, ses recherches se sont portées sur la collaboration homme-robot et l'intérêt de la réalité virtuelle pour la conception de postes collaboratifs. Depuis 2015, Vincent Weistroffer est ingénieur-chercheur au CEA-List (laboratoire d'intégration de systèmes et des technologies), dans le laboratoire de simulation interactive (LSI), et anime une équipe autour de l'interaction en environnement virtuel. Ses centres d'intérêt sont variés et concernent la réalité virtuelle, la simulation de gammes de montage, la capture de mouvements et la simulation de mannequin biomécanique, notamment pour l'analyse ergonomique des postes de travail.
→ ***La « cobotique » et l'interaction homme-robot***