FAUT-IL CRAINDRE L'INTELLIGENCE ARTIFICIELLE ?

Benoit Le Blanc

Professeur à Bordeaux INP, spécialiste en Intelligence Artificielle et en Sciences Cognitives, directeur de l'Ecole Nationale Supérieure de Cognitique (ENSC Bordeaux INP).

Faut-il craindre l'Intelligence Artificielle (IA) ? Depuis plusieurs dizaines d'années, cette question se pose très régulièrement. Elle revient systématiquement dès qu'est présentée une nouvelle avancée technologique bluffante. Elle agite toujours les mêmes ressorts émotionnels de la peur, de l'angoisse, du malaise. A chaque fois, la réponse prudente et raisonnée est inaudible ou très vite oubliée. A croire que décharges d'adrénaline et poussées technologiques doivent aller de pair. Aussi disons le tout de suite : non, il ne faut pas craindre l'IA. En revanche il nous faut bien reconnaître qu'il va devenir indispensable de nous méfier de celles et de ceux qui en feront une utilisation malveillante, autant que de celles et de ceux qui en font déjà un usage inconsidéré, mal compris ou à leur seul profit. Pour aborder cette question, nous proposons une étude en trois points : nous posons tout d'abord quelques principes permettant de comprendre ce qu'est l'IA ; nous illustrons ensuite ce qui se fait avec l'IA ; et nous réfléchissons enfin à comment avancer socialement et politiquement avec l'IA.

CE QU'EST L'INTELLIGENCE ARTIFICIELLE

Depuis toujours, l'IA s'est inscrite, et continue à s'inscrire, dans le processus d'informatisation. **Les pionnières et les pionniers** du domaine sont aussi celles et ceux qui ont construit l'informatique moderne et le calcul automatique. Pas d'IA sans informatique, certes, mais en retour pas de science des ordinateurs sans une réflexion sur ce que peut être et ce que peut faire l'IA.

L'informatisation passe par la numérisation, c'est-à-dire par des étapes de *réduction*, d'*approximation* et finalement de *codage*. Avant de devenir des données, les informations du monde qui nous entoure doivent consentir à une réduction de ce monde. Aborder la réalité se fait en y plaquant une grille, là où l'on pense qu'une information pertinente se situe. Notons déjà qu'au delà du périmètre de cette grille, aucune donnée ne prendra forme, sauf à appliquer une autre grille à coté de la première. Le monde est alors *réduit au périmètre* des grilles de numérisation. A l'étape suivante, chaque case de la grille se verra attribuée une sorte d'étiquette uniforme, venant ainsi *approximer toutes les nuances* que chaque case pouvait comporter. Ces étiquettes ou désignations ont un caractère arbitraire certain, mais participent à une étape nécessaire pour pouvoir obtenir une donnée. Vient enfin le *processus de codage* qui définit les seuls états qui seront pris en considération et donc les seules valeurs que l'on pourra attribuer à chacune des cases de la grille que l'on considère. Tout ce qui est hors de la grille, tout ce qui est autre que l'étiquette de chaque case, ou bien tout ce qui est plus fin que la différence entre deux valeurs d'étiquettes, n'existe tout simplement pas dans le monde numérique.

Voila dressée en quelques mots ce qui constitue la base de l'informatisation, et donc également celle de l'IA. L'information doit rentrer dans un format pour devenir une donnée. Alors que chez l'humain, l'information doit pouvoir être plongée dans un contexte sémantique pour devenir une connaissance. On a bien là deux modes de fonctionnement et deux mondes différents : le **nôtre**, dans lequel informations et contextes composent les connaissances ; et le monde des machines, dans lequel informations et formats ouvrent à la capacité de stockage et de traitement des données. Certains voient là une sorte de continuum : données, informations, connaissances, savoir. C'est avec cette vision que l'on peut développer par exemple le domaine de la gestion des connaissances, en œuvrant pour formaliser les connaissances humaines et capitaliser le savoir des experts. D'autres diront que les connaissances sont et restent une exclusivité humaine, en conséquence de quoi les machines ne peuvent y accéder et se contenteront de ne gérer que des données... Pour les premiers, l'intelligence de l'IA peut se construire en capturant l'intelligence humaine ; pour les second, il est illusoire de parler d'intelligence dans le cadre de l'IA.

En plus d'avoir ses deux pieds bien ancrés dans l'informatique, l'IA a

l'ambition de se comparer à l'intelligence humaine, au cerveau animal et aux neurones biologiques. Mais sait-on véritablement comment toute cette matière vivante fonctionne ? On cherche, on propose, on conjecture et, à chaque avancée, ce sont de nouvelles idées qui viennent fertiliser l'IA, de nouveaux espoirs qui sont amplifiés dans les médias, de nouvelles sirènes qui retentissent chez les collapsologues. Pour s'y retrouver, il nous faut poser quelques points de repère. Tout d'abord l'intelligence humaine est une chose, l'intelligence artificielle en est une autre. Dans leurs principes, leurs constructions et leurs évaluations, ces deux choses sont différentes. Bien sûr l'IA va s'inspirer des découvertes sur le cerveau, le corps ou le comportement, mais l'IA gardera les limites inhérentes à l'informatique. Même si l'on imagine ce que pourrait être un esprit pur, et que l'on cherche à y tendre en souhaitant des cerveaux informatiques, téléchargeables, et qui puissent être transférés, en retour il nous faut admettre que les seuls cerveaux que nous voyons dans la nature sont humides, dans des corps et **qu'on les** voit vieillir. Rien de tout cela n'est modélisé en IA... Par ailleurs, il est amusant de remarquer que les avancées en IA poussent les neurologues, les physiologistes et les psychologues à reconsidérer leurs points de vue, sur leurs propres modèles et leurs propres recherche en faisant usage du vocabulaire et des métaphores de l'IA. On parle ainsi de mémoire à court terme et de mémoire à long terme chez l'humain, alors que cette distinction n'a été posée en informatique que pour des questions de taille des zones consacrées aux registres de stockage.

Pour faire simple, au cours de son histoire l'IA s'est structurée et continue à avancer selon deux voies différentes : celle des symboles et celle des configurations. Ces deux concepts viennent beaucoup plus traduire l'état d'esprit des différents chercheurs en IA qu'une nécessité physique ou matérielle. Cette forme de dualité existe aussi par exemple entre l'Histoire et la Géographie, entre la Psychologie et la Sociologie, entre l'Electronique et l'Informatique, etc. A chaque fois, il s'agit de deux points de vue portés sur un même concept : celui du territoire, de l'individu, du calcul, etc. Pour l'IA c'est le concept de prise de décision qui constitue le centre des préoccupations : comment permettre à un programme ou un automatisme de prendre une décision, ouvrant à une action. Chronologiquement ces deux voies de l'IA sont arrivées en même temps, **c'est-à-dire** dès le début de l'IA, et ont ensuite chacune tour à tour servi des intérêts stratégiques et économiques, avant de laisser le devant de la scène à l'autre voie. A chaque étape, les avancées technologiques ont constitué de véritables bons en avant. On peut citer le Perceptron, le système expert Mycin, le programme de jeu d'échec Deeper Blue, le Knowledge Graph de Google, le programme Alphago de DeepMind, le projet Debater d'IBM, l'agent conversationnel ChatGPT d'OpenAI, etc. Ces avancées sont de plus en plus rapides et se combinent dès lors qu'il s'agit de composer des produits commerciaux, comme pour les identifications faciales sur les photos telles que celles proposées par Facebook, les recommandations de produits telles que celles proposées par Netflix ou Amazon, les recherches d'informations telles que celles proposées par MicroSoft ou Google, etc.

La voie de l'IA qui explore les symboles est appelée le cognitivisme. L'intelligence de la machine reposerait alors sur la capacité à manipuler ces symboles, un peu comme le langage nous permet d'assembler des mots et in fine nous permet de faire passer une idée. Les premières réalisations ont consisté à proposer une modélisation du monde à explorer sous la forme d'un graphe et d'y coller des algorithmes astucieux de parcours. Pour développer ces algorithmes, des langages de programmation tels que Prolog ou Lisp ont été mis au point. L'idée générale consiste à enchainer des règles et à manipuler les formes de connaissances que l'on aura extraites de l'expertise d'un humain. Ce passage d'une connaissance implicite d'un expert, à des éléments explicites manipulables par une machine, est à la source du cognitivisme. L'intelligence reste humaine et se retrouve simulée sur une machine par le jeu d'enchainement de symboles. Les perspectives sont de pouvoir mettre le maximum de symboles dans une machine pour obtenir une IA performante.

La voie de l'IA qui explore les configurations est appelée le connexionnisme. L'intelligence de la machine émergerait alors de la capacité à connecter des configurations. On parle d'une approche sub-symbolique car ce sont les configurations qui vont ici jouer le rôle qu'ont les symboles dans le cognitivisme précédent. Sans que ces configurations ne puissent réellement correspondre à quelque chose d'identifiable pour un humain, elles vont servir de base aux calculs de la machine. La reconnaissance de motifs dans des images, des sons, ou tout autre type de signaux, ouvre la voie à des opérations de classification et d'enclenchement d'actions dédiées. Pour ce dispositif de reconnaissance de motifs appris, ce qui est codé dans la machine correspond à ce que l'on pense être du processus de caractérisation de l'intelligence, de la reconnaissance immédiate faite par l'oeil humain. Les perspectives sont de pouvoir réduire au minimum la taille de ces motifs pour obtenir une IA performante.

Précisons que dans un cas comme dans l'autre, les informaticiens parlent de *l'hypothèse du monde clos* pour expliquer doctement que tout ce qui n'est pas codé dans la machine sera considéré comme faux ou conduira à des résultats inconnus. Votre dispositif d'IA enchaine les règles édictées par un expert automobile ? A vous de savoir si cet expert est un motoriste, un agent de la circulation ou un assureur... car tout emploi de ce dispositif en dehors de son cas d'usage ne sera pas nécessairement bien adapté. Votre dispositif d'IA reconnaît les vingt-six lettres de l'alphabet ? Et bien on ne sait absolument pas ce qu'il produira face à un symbole tel que celui de l'euro... Et d'ailleurs avez vous interrogé aussi bien des droitiers que des gauchers pour produire les motifs des lettres manuscrites venant alimenter l'apprentissage de la machine ? Toute confrontation à une situation inconnue par la machine entrainera un comportement imprévisible, bien que choisi dans la panoplie des actions prédéterminées pour cette machine.

Au cours de l'histoire de l'IA, la discipline a suivi une évolution que l'on peut

retracer à travers les définitions qui ont été formulées au fil du temps pour décrire ce domaine scientifique. On peut placer un début de cette histoire avec l'article d'Alan Turing de 1950 dans lequel il ne parle ni d'ordinateur (les « *computeurs* » désignaient à l'époque les personnes qui effectuaient les calculs), ni d'IA (le terme n'a été proposé que deux ans après le décès de Turing) mais ce sont portant bien ces travaux de Turing qui ont permis de décider du déchiffrage des codes de communication allemands pendant la seconde guerre mondiale, et en ce sens constituent une première réalisation d'IA.

La définition initiale de l'IA est ambitieuse, elle est associée à un cycle de conférences durant l'été 1956 : « *Chaque aspect de l'apprentissage ou toute autre caractéristique de l'intelligence peut en principe être décrit si précisément qu'il est possible de construire une machine pour le simuler* ». Les auteurs d'alors avaient en tête le jeu de l'imitation proposé par Turing. Ce test de Turing pointait le fait que dans quelques dizaines d'années, suite à n'importe quelle question formulée, il ne serait plus possible de distinguer si les réponses sont faites par un humain ou bien par une machine.

Dans les années 1970, des chercheurs comme Margaret Boden ou Marvin Minsky utilisaient une désignation plus pragmatique et finalement moins audacieuse : « *Le but de l'IA est de construire des machines qui réalisent des choses qui requièrent de l'intelligence lorsqu'elles sont faites par des humains* ». En 2017, à l'occasion du programme inter-ministériel #FranceIA et étant à cette époque en responsabilité du domaine de l'IA au sein de la DGRI, nous avons produit une définition plus large et interdisciplinaire : « *Aujourd'hui, l'intelligence artificielle désigne un ensemble de notions ou d'algorithmes, s'inspirant de la cognition humaine ou du cerveau biologique, et destiné à assister ou suppléer l'individu dans le traitement des informations massives* ». L'idée était de porter en évidence, d'une part que désormais le champ de l'IA dépassait largement celui de l'informatique, et d'autre part que l'économie venait jouer un rôle majeur dans le développement du domaine. Ce sont certainement ces idées qui ont été reprises par la Commission d'enrichissement de la langue française quand elle a publiée au Bulletin Officiel de la République Française du 9 décembre 2018 cette nouvelle définition (la précédente datait de 1989) : « *Champ interdisciplinaire théorique et pratique qui a pour objet la compréhension de mécanismes de la cognition et de la réflexion, et leur imitation par un dispositif matériel et logiciel, à des fins d'assistance ou de substitution à des activités humaines* ».

CE QUE L'ON FAIT AVEC L'INTELLIGENCE ARTIFICIELLE

Reconnaitre des visages sur une photo, transformer un texte manuscrit en un fichier texte, traduire ce texte, déterminer sur une image si un véhicule est un vélo, une voiture ou un camion par exemple, suivre le parcours de ce véhicule en mouvement sur une vidéo, lire sa plaque d'immatriculation s'il en a une, proposer un morceau de musique à écouter, ou encore déterminer un itinéraire, sont autant d'actions qui composent notre quotidien et qui, lorsqu'elles sont automatisées, utilisent abondamment les résultats de la recherche en IA. Pour chacune d'entre elles, il s'agit de mettre en oeuvre une forme de prise de décision, par le classement du signal reçu, dans une catégorie apprise. La performance du système dépend donc directement de la quantité de catégories envisagées à l'avance et de la représentativité de ces catégories pour faire face aux signaux qui seront reçus par le système. Car tous ces systèmes fonctionnent en deux temps. Le premier temps est celui du calibrage, de l'initialisation des valeurs de façon à assurer une bonne reconnaissance d'un ensemble déterminé de formes présentées au système. Puis le second temps est celui de l'utilisation, en proposant les signaux en situation réelle et en s'appuyant sur la classification qu'en fait le système.

A coté de cela s'est développé depuis 2016 un nouveau domaine : celui de l'IA générative. Il ne s'agit plus seulement de reconnaître une catégorie apprise, mais d'utiliser ces catégories pour générer un signal vraisemblable. Cette nouvelle approche revisite complètement la manière de travailler des publicitaires ou des cinéastes. Elle ouvre aussi une nouvelle voie dans les capacités de traduction des textes et de création d'applications autour des *agents conversationnels*. Derrière cette IA générative se cachent trois concepts : l'apprentissage profond, les réseaux antagonistes et les modèles de langage...

La méthode algorithmique la plus utilisée actuellement en IA est celle des réseaux de neurones formels, plus particulièrement celle des réseaux à propagation en couche, et plus précisément celle des modèles d'apprentissage profond. On le voit bien dans cette phrase, les informaticiens ont la fâcheuse tendance à nommer ce qu'ils font à partir de leur propre vocabulaire, et donc de s'approprier certains mots courants, en les détournant de leur sens usuel, pour leur donner un sens spécifique à l'informatique. **Comprenne** qui pourra ! Bien évidemment, il n'y a pas de neurones dans une machine, mais il y a des structures de données sous la forme d'automates à seuil de déclenchement ; il n'y a pas non plus de véritable capacité d'apprentissage par la machine, mais il y a des structures de contrôle positionnant des exemples pré-déterminés sur les valeurs propres d'une matrice de transition d'un espace vectoriel de très, très, très grande dimension. Quoiqu'il en soit, le modèle des réseaux de neurones formels vient proposer une solution à l'automatisation d'un mécanisme associatif entre un signal d'entrée et une catégorie de sortie, même lorsque cette association ne se

fait pas de façon linéaire : c'est-à-dire lorsque l'on ne peut pas séparer par un simple trait, les différentes catégories recherchées dans les signaux. Et c'est bien évidemment le cas le plus courant ; on parle de problèmes non linéairement séparables.

Les bases de l'apprentissage profond, ou Deep Learning (DL), ont été proposées dès les années 80 (notamment par Y. Le Cun), mais le DL nécessite beaucoup de calculs pour pouvoir atteindre une performance rivalisant avec d'autres types de solutions. Ce n'est donc qu'en 2012 qu'un programme de DL a pu dépasser les autres solutions dans le domaine de la reconnaissance d'images. Le terme « profond » provient du fait que plusieurs couches successives de groupes d'automates viennent traiter et transformer les informations initiales. On parle de couches de neurones convolutives car l'opération de convolution y joue un rôle important. Il s'agit d'appliquer une petite fenêtre de quelques pixels (typiquement 3x3 ou 5x5) sur l'information délivrée par la couche de neurones précédente et d'en tirer une valeur globale pour cette petite fenêtre. L'opération est reproduite de proche en proche, sur toute la surface de l'image, et c'est donc une nouvelle image qui est ainsi fabriquée, en traduction des continuités ou ruptures présentes initialement. Ce processus de convolution est remis en œuvre sur une succession de couches, en étant entrecoupé de couches de traitement plus classique. A la fin, la machine a déterminé grâce à cette suite de calcul, une répartition des informations à apprendre dans un espace assez grand pour les loger et les distinguer toutes. Le réseau est alors prêt à être utilisé, en lui fournissant une information nouvelle pour lui, qui va être classée dans l'une des catégories apprises.

Si vous disposez par exemple d'une dizaine de photos de visage pour chaque individu dans une population d'un million de personnes (cela fait déjà dix millions de photos), vous pouvez composer un réseau de neurones destiné à reconnaître chacune de ces personnes (il y a un million de catégories en sortie, soit une pour chaque individu). Quand quelqu'un se présente ensuite devant une caméra de surveillance par exemple (il s'agit donc d'une nouvelle photo, jamais encore vue par le système), il est alors possible de déterminer qui est cette personne parmi le million d'individus connus. Le système doit être assez robuste pour fonctionner même si la personne porte de nouvelles lunettes ou se coiffe différemment car les informations codées ne sont pas les informations sur le nez, les oreilles, ou sur tout autre signe que nous autres humains jugeons distinctif. Les informations stockées sont propres au fonctionnement de la machine et dépendantes de l'ensemble des visages-catégories. Si vous portez un masque représentant le visage de quelqu'un de précis, il n'est même pas certain que le système vous associe à cette personne ! Ce qui a du sens pour nous, n'est pas nécessairement le point d'ancrage des repères de la machines, et en retour, ce qui constitue les valeurs d'apprentissage de la machine, n'a vraisemblablement aucun sens pour nous.

En 2016 un chercheur (I. Goodfellow) a proposé d'utiliser un modèle d'IA composé de deux réseaux de neurones, pour produire des visages fictifs. Son idée a été de démarrer avec un réseau de neurones qui fonctionnerait « à l'envers ». Plutôt que de partir d'une image et d'en trouver la catégorie, le premier réseau part de catégories et propose un visage par son calcul. Bien évidemment ce visage fabriqué n'en sera pas un et ne ressemblera à rien du tout, car il ne sera composé que de pixels répartis au hasard. Vient alors un second réseau de neurones préparé en entrée avec un ensemble d'images dont certaines sont de vrais visages, et disposant de seulement deux catégories de sorties : « visage » ou « non-visage ». Ce second réseau permet de juger si la production du premier est crédible ou non. Si le premier réseau produit n'importe quoi, il doit alors revoir ses valeurs d'attribution (appelées poids des neurones) et recommencer. Les deux réseaux sont dits « antagonistes » et on parle de Generative Adversarial Networks (GAN). C'est la combinaison de ces deux réseaux de neurones, l'un génératif, l'autre examinateur, qui ouvre la voie à l'IA générative, en capacité de produire des « fake faces », « fake videos » et toutes autres images irréelles.

Le principe de la convolution qui est à la base des calculs de ces réseaux fait que les pixels sont générés par petites fenêtres, de proche en proche, et qu'il est donc impossible pour la machine d'établir un véritable lien de symétrie entre les différentes parties de l'image produite. Ceci explique que **les visages générés** ont parfois des asymétries gauche/droite, ce qui n'est pas gênant car c'est souvent le cas dans la nature. En revanche les objets comme les chapeaux ou les paires de lunettes ont aussi des asymétries qui, elles, sont troublantes. Ces programmes d'IA générative doivent alors ajouter des traitements pour redresser ces éléments qui nous permettraient sinon d'identifier immédiatement la supercherie.

Lorsque la base d'apprentissage du GAN ne concerne plus des visages, ni même des images, mais le langage, il doit alors disposer d'un très grand modèle de ce langage. On parle de LLM, pour Large Language Models. C'est à partir de 2018 que ces modèles ont été mis au point, dans l'objectif de fabriquer des agents conversationnels. Ils se basent sur des textes connus et y déterminent toute une série de concordances entre les mots : après chaque mot, la probabilité d'apparition de tel ou tel autre mot **est de telle ou telle valeur**, etc. Si trois mots avant, il s'agissait de tel mot, alors cette probabilité se modifie, etc. Le modèle du langage est ainsi fabriqué par des probabilités conditionnelles sur les occurrences de mots, de groupes de mots, voir de phrases entières. Les agents conversationnels basés sur ces modèles de langage vont alors générer un texte, mot à mot, non pas en fonction de connaissances particulières, mais en fonction de probabilités conditionnées par toute une série de paramètres. Et cela fonctionne ! Le plus bluffant avec ce dispositif est que l'on obtient un texte crédible, écrit tout à fait correctement, dans une langue que l'on peut choisir, et si on le souhaite répondant à des paramètres initiaux, tels que « en deux lignes », ou bien « à la manière de Victor Hugo », ou encore « dans un ton colérique », etc.

L'IA générative produit des éléments (textes, images, vidéos, et bientôt sons) plus souvent faux que vrais, car il **n'est possible**, au hasard des paramètres, de tomber sur quelque chose de vrai que si les enchainements des mots correspondent à des locutions dédiées à quelque chose de précis. Les opérateurs d'IA générative cherchent bien évidemment à renforcer cette véracité de la production, en jouant sur des règles de production pilotées et contrôlées par des humains. Quoiqu'il en soit les domaines très techniques ou faibles en signification (comme par exemple la production de codes informatiques, de slogans publicitaires, de scénarios basiques, de résumés de texte) vont très rapidement être envahis par des productions issues d'IA générative.

La plupart des gens associent les productions de Dall-E ou de Midjourney à des « inventions », ou des « créations ». Le vocabulaire employé oublie que ces machines n'ont pas de conscience, pas d'intention, pas de jugement. Cet anthropomorphisme qui prête aux machines des opinions, de la politesse, des idées, est très gênant car il témoigne d'une sorte de crédulité spontanée, pouvant devenir une porte d'entrée à la manipulation des esprits. Par exemple, le terme de « hallucination » a commencé à être employé à partir de mars 2023, pour désigner des productions de ChatGPT crédibles mais invraisemblables. Il ne s'agit pourtant là ni d'une procédure particulière du programme, ni d'une dérive de ce programme. Dans son fonctionnement normal, ChatGPT génère des phrases crédibles mais n'a pas d'organe des sens à la base de cette composition. L'emploi d'un nouveau terme anthropomorphe vient un peu plus brouiller les pistes.

CE QUI VA CHANGER LES CHOSES AVEC L'IA

L'intelligence artificielle pénètre de plus en plus le fonctionnement des technologies et des outils actuels. En automatisant certaines reconnaissances de formes ou de configurations, le but est d'enclencher des actions adaptées à ces reconnaissances. Parmi ces actions adaptées, il y a la production de nouveaux contenus, générés automatiquement. Les gens vont utiliser l'IA générative pour obtenir des textes plausibles, demander quelles sont les principales idées présentes dans ces textes, quelles sont les références associées à ces idées, obtenir du système ces références... et ne pas se rendre compte que tout cela a été généré, donc se présente sous la forme de quelque chose **de** vraisemblable, mais pas **de** nécessairement vrai. La plupart du temps ces références n'en sont pas, ou plus exactement peuvent être des références, mais pour d'autres sujets.

Un avocat new-yorkais en a subi les conséquences. Dans sa préparation d'un procès en début d'année 2023, il a utilisé ChatGPT pour construire son argumentation. Sans s'en rendre compte, il a ainsi généré une jurisprudence totalement fictive, en s'appuyant sur des références crédibles mais imaginaires. Le juge l'a confondu et cette affaire a été relatée dans la presse mondiale.

Avocats, publicitaires, journalistes, scénaristes, enseignants : tous ces métiers vont très vite être confrontés à l'usage de l'IA générative. Il y a dix ans, une étude avait déjà mis en avant le problème de la mutation ou de la disparition des métiers, du fait de l'informatisation et de l'IA devenant opérationnelle. Les auteurs de cette étude (C. Benedikt et M. Osborne) ont essayé de quantifier la part informatisable de chacun des métiers et en ont extrapolé que près de la moitié des métiers actuels disparaitront ou seront bouleversés par l'informatique. Il s'agit là d'une situation alarmiste, forçant un peu le trait, car ne prenant pas en compte la temporalité des évolutions dans les professions.

Pourtant, un processus de mutation est certainement en œuvre. Si l'on se réfère à l'informatisation apparue dans les années 1980 et accélérée dans les deux décennies suivantes, on a pu voir un passage progressif du papier aux fichiers, et l'apparition d'écrans sur à peu près tous les outils. Des entreprises se sont créées en commençant par poser leur système d'information et en structurant les différents métiers autour de cet axe. A l'heure actuelle, ce sont les données massives qui tirent les innovations en entreprise. Tout est numérisé, on a abandonné la trace papier. On est passé du rapport écrit au tutoriel : ce n'est plus la documentation qui est recherchée mais plutôt des exemples filmés de situations qui lui sont préférés. Ces vidéos que l'on retrouve maintenant un peu partout, et qui composent les supports de formation, vont bientôt constituer les points d'ancrage des gestes professionnels. Dans nos années 2020, un schéma d'articulation se met en place à partir des données massives, des bases de vidéos et des algorithmes d'IA venant recommander la vidéo présentant le bon geste à accomplir en fonction de la situation reconnue selon les données captées. La notion de métier devient un peu plus floue et la reconfiguration des gestes va venir structurer l'exercice même du travail.

La taille gigantesque de ces bases de données dépasse l'entendement humain. On parle de données massives pour désigner des informations en très grand nombre, en très grande variabilité de formes et en très grand flux de production. Jusqu'à l'ordre de grandeur du milliard, les gens peuvent se faire une représentation mentale utile de la chose. Au delà, la quantité est vue comme un infini, et toute croissance supplémentaire devient très difficile à percevoir. C'est pourquoi l'exponentielle n'est pas une fonction facile à manipuler. Même si l'histoire est fausse, on raconte que l'inventeur du jeu d'échec avait demandé à son Prince, séduit par ce jeu, une récompense en grains de blé : 1 grain sur la première case, 2 sur la deuxième, 4 sur la troisième, 8 sur la suivante et ainsi de suite en multipliant par deux à chaque case. Avant d'avoir atteint la moitié de l'échiquier, on a déjà plus d'un milliard de grains par case, et bien évidemment la croissance exponentielle continue. Si tout un chacun peut se faire des repères pour comprendre comment **l'augmentation a lieu** sur la première moitié de l'échiquier, plus personne n'en a pour comprendre la suite de cette croissance. Les potentialités liées aux calculs sur les machines croissent de façon

exponentielle depuis déjà des dizaines d'années... aussi les nouveautés arrivent maintenant plus vite que l'on ne peut les assimiler.

Si les individus ne peuvent plus se faire une idée juste des capacités accrues des machines et des potentialités offertes par l'IA, il devient alors du ressort des dirigeants **politiques de se saisir** de ce sujet. C'est ce qu'a fait l'administration de B. Obama en 2016 en rendant public un rapport intitulé *Preparing For the Future of AI - Final Report from National Science and Technology Council.* Dans les deux années qui ont suivies, chaque pays du monde s'est doté d'une stratégie en IA, de façon à consolider ou étendre les bases de sa recherche scientifique, et aider administrations et entreprises à se constituer de larges bases de données exploitables par les algorithmes d'IA.

En France la stratégie IA a été dévoilée par le Président de la République en 2018. Un troisième volet a été ajouté, en rapport avec la réflexion sur les enjeux éthiques de l'IA. Travaillant alors au Ministère de la Recherche, pour le domaine scientifique de l'IA, j'ai été amené à discuter avec nos homologues allemands, canadiens et japonais de ce sujet. Tous étaient conscients des enjeux et soucieux d'agir au niveau politique sur ce thème, anticipant le fait que l'IA viendra très vite changer les relations entre individus. L'idée de construire un « GIEC de l'IA » a été rapidement écartée, car il ne s'agit pas d'étudier un domaine comme peut l'être le climat, c'est-à-dire un sujet dont on ne connait pas les contours et les règles de fonctionnements. Ici, il s'agit plutôt d'étudier les possibilités offertes par une IA que les scientifiques maitrisent, en tant qu'objet technique. Le problème avec l'IA est d'arriver à donner un cadre aux agissements des individus. **On n'en fait** pas un « GIEC de l'énergie » ou un « GIEC de l'économie ». La structure adoptée a donc été celle d'un GPAI (Global Partnership for AI) et a été confiée à l'OCDE.

Dans la mise en place de ce GPAI, il m'a été demandé de dresser quelques axes pour structurer les actions d'un tel groupe de réflexion. Ma proposition a été de porter l'attention sur trois niveaux : celui de l'interaction avec les machines (le niveau de l'algorithme), celui relatif aux données personnelles (le niveau de l'individu), et celui concerné par des mises en action variées comme dans le commerce par exemple (le niveau de la société). A cela a été ajouté un quatrième niveau pour traiter des relations internationales (le GPAI est une réflexion entre plusieurs pays).

Les thèmes proposés alors pour le mandat de ce GPAI concernaient la collecte et **l'équilibre** des données ; le contrôle des données et le respect de la vie privée ; le partage des modèles et des connaissances ; la validation, la qualification, l'intelligibilité et la transparence des techniques d'IA ; l'acceptabilité et l'appropriation de l'IA ; l'avenir du travail ; la gouvernance, les lois et la justice face à l'IA ; l'IA responsable et les droits de la personne ; et enfin l'équité, la responsabilité et les biens publics. Tous ces thèmes sont génériques mais leur

simple liste montre bien que des pans entiers de notre société vont être modifiés par le fait que des prises de décision automatiques sont rendues possibles par l'IA.

Une autre manière d'aborder les questions de l'éthique en IA peut s'exprimer en dressant deux axes de réflexion. L'un de ces axes vient décrire le volet proximal[1] ou bien distal[2] de la décision ; l'autre axe vient qualifier la manière dont les données sont recueillies, par déclaration consentie ou bien par trace des activités épiées. Plus orientée vers la personne et moins dépendant de la technique, cette approche est issue de discussions tenues avec des collègues professeurs en sciences humaines (notamment J. Dubucs[3] et B. Claverie[4]). Le premier axe, celui de la prise de décision, peut se voir comme un axe reliant la sociologie à la psychologie ; l'étude de l'action distale va bénéficier des sciences de la société, alors que l'étude de l'action proximale va bénéficier des sciences du comportement. Le second axe, celui des données, peut se voir comme un axe reliant les besoins en cadrage politique ; il y a un besoin d'éducation des personnes pour les avertir **quant aux** les conséquences liées aux données qu'elles déclarent, et il y a un besoin de protection des personnes en obligeant à « **anonymiser** » les données relevées automatiquement dans les traces de visite des sites web ou dans les relevés d'usages des applications sur téléphones mobiles.

Une autre source d'inspiration intéressante pour penser l'éthique de l'IA est à prendre dans les délibérations du Conseil d'Etat. Lors de la révision des lois sur la bioéthique en 2018, j'ai participé à certaines séances de discussion pour le compte de la DGRI, notamment pour traiter de la question de l'incidence de l'IA et des données numériques sur la bioéthique. Cette question de l'IA et des données sera certainement au menu des prochaines révisions qui interviennent tous les sept ans, mais ce sujet n'était pas encore mûr au tout début des années 2020. En revanche l'une des principales avancées de cette révision a été en 2021 l'autorisation de la procréation médicalement assistée (PMA) pour toutes les femmes, et le maintien de l'interdiction de la gestation pour autrui (GPA). L'explication de cette asymétrie dans l'autorisation ou non d'enfanter pour des couples homoparentaux tient dans la ligne de conduite du Conseil d'Etat : dignité, solidarité, liberté. Bien avant l'enfant, c'est la femme et sa dignité qui **guident** le débat sur la procréation. Ainsi, avant même la question de laisser libre de faire ce qui est techniquement possible, se pose la question de la solidarité et donc de la capacité à proposer cette action à tous et à toutes, et encore avant, donc en tout premier lieu, se pose la question de la dignité de la ou des personnes impliquées dans la réalisation de cette action que la technologie rend possible.

---

[1] Fait référence aux éléments immédiats ou aux conséquences directes d'une décision

[2] Englobe les résultats ou les incidences qui se manifestent après un certain laps de temps

[3]

[4]

Employé dans le cadre de la bioéthique, ce triptyque pensé par le Conseil d'Etat a été plusieurs fois repris dans les débat à l'Assemblée Nationale, comme on peut le lire dans le Journal Officiel de la République Française du 8 juin 2021. La dignité d'abord, c'est-à-dire la primauté de la personne humaine, le respect de l'être humain dès le commencement de sa vie, l'absence de caractère patrimonial du corps humain, l'inviolabilité et l'intégrité de l'espèce humaine ; la solidarité ensuite, qu'illustre l'altruisme des donneurs et oblige à entendre la souffrance exprimée par celles et ceux qui sont confrontés à un besoin : en santé, la solidarité se traduit notamment par l'assurance maladie obligatoire et l'égal accès aux soins ; la liberté, enfin, qui vise à préserver la part de vie privée et donc l'autonomie de l'individu dans ses choix les plus intimes.

Dignité, solidarité, liberté : voilà le triptyque sur lequel repose notre modèle de bioéthique depuis plus de vingt-cinq ans et dont on pourrait s'inspirer pour élaborer une éthique de l'IA. La dignité, c'est-à-dire le respect que mérite quelqu'un, se traduit par une protection particulière de la personne face à une technologie et un monde numérique qui vient bouleverser nos rapports des uns envers les autres. La solidarité, se retrouve dans la conception du don altruiste et dans l'attention portée aux plus vulnérables : vulnérables car moins conscients de ce qu'est ce monde numérique et moins en capacité d'y agir car se trouvant en déficit de connaissance face à des entreprises dont la place économique se fait en monétisant les données des personnes. La liberté individuelle, s'exprime enfin dans l'équilibre personnel entre éléments consentis et bénéfices tirés de ces prises de décision automatisées grâce aux algorithmes d'IA.

Remarquons que les deux axes cités plus haut (axe sur la portée des décisions, croisé avec l'axe sur les données) constituent une piste pour tracer la façon dont pourrait être définie la dignité de la personne dans un monde numérique.

POUR CONCLURE

Les travaux sur l'éthique de l'IA sont nombreux, beaucoup de voix se font entendre et plusieurs arènes de discussion se montent, mais je n'ai encore vu nulle part une orientation qui tirerait profit de ce triptyque du Conseil d'Etat. Pourtant le temps passe, les progrès techniques s'amoncèlent et la réponse politique doit se structurer pour que dignité, solidarité et liberté ne **laissent** pas la place à honte, pitié et exploitation.