Comprendre ou transformer l'humain ? Enjeux et voies potentielles de l'Intelligence Artificielle

Frédéric Alexandre

Comprendre ou transformer l'humain ? Enjeux et voies potentielles de l'Intelligence Artificielle

Frédéric Alexandre, Inria Bordeaux Sud-Ouest, Laoratoire Bordelais de recherche en informatique UMR 5800, Institut des Maladies Neurodégénératives UMR 5293

Introduction :

Les pistes de réflexion que je propose ici reposent sur une expérience de plus de trente ans, que j'ai acquise dans la plupart des domaines de l'Intelligence Artificielle que je vais évoquer ici, dans ses aspects numériques et symboliques et dans ses approches faibles et fortes. Comme je le développerai en conclusion de ce texte, c'est aussi cette expérience qui m'a amené à orienter ma trajectoire vers les neurosciences cognitives et la médecine, en rejoignant un laboratoire de neurosciences sur le campus de l'Hôpital de Bordeaux, et à m'intéresser également à l'intelligence naturelle ainsi qu'au cerveau et à ses dysfonctionnements. Ce parcours me place dans une situation privilégiée pour observer et analyser les développements de l'Intelligence Artificielle et la manière dont ils viennent modifier notre société et en premier lieu influencer nos cerveaux. On nous parle de plus en plus souvent de technologies intelligentes qui vont nous rendre la vie plus facile en nous aidant dans notre quotidien et dans notre vie professionnelle ; on s'alarme tout aussi régulièrement du fait que ces mêmes technologies pourraient nous asservir et même nous remplacer. Je propose d'aborder de telles questions en ayant une vue informée de la situation.

Par les changements annoncés dans notre société, les technologies de l'Intelligence Artificielle nous promettent le meilleur ou le pire. Par leur nombre croissant, elles nous promettent en tout cas un impact de plus en plus large et il est donc important de s'y préparer et, le cas échéant, de se protéger de certains effets indésirables. Pour cela, il est tout d'abord important de savoir définir l'Intelligence Artificielle, concernant ses principes généraux aussi bien que certaines caractéristiques plus techniques, afin d'évoquer ses buts et ses niveaux de performance associés. En effet, certains jugements que l'on porte à son propos peuvent être dûs à une mauvaise connaissance de ce qu'est aujourd'hui l'Intelligence Artificielle et de ce qu'elle prétend réaliser. Nous nous attacherons donc à commencer à clarifier ces aspects dans une première partie de ce texte.

Il est probable qu'une partie des erreurs de jugement que je viens d'évoquer est due au fait que l'Intelligence Artificielle se réfère au phénomène de l'intelligence, tel qu'on peut l'observer chez les animaux et les humains et que l'on puisse donc avoir une tendance à généraliser entre ces différentes formes d'intelligence, artificielle et naturelle. Pour aider à tracer la frontière, j'évoquerai dans une deuxième partie l'intelligence telle qu'elle est produite par le cerveau d'un être vivant et telle qu'elle est aujourd'hui étudiée par les neurosciences cognitives. En contraste, ceci me permettra de souligner les différences entre intelligences artificielle et naturelle, en se demandant toutefois si ces différences ont vocation à s'estomper avec l'amélioration des technologies numériques ou s'il y a des domaines où l'Intelligence Artificielle n'a pas vocation à se déployer.

Cette analyse pourra servir de base pour lancer la discussion proposée dans la troisième partie de ce texte. Dans l'état actuel de ces technologies, quels sont les risques associés à l'utilisation

de l'Intelligence Artificielle, mais aussi quels sont les nombreux fantasmes couramment véhiculés sur l'Intelligence Artificielle, qui n'ont pas de base vraiment sérieuse aujourd'hui et nous détournent peut-être des risques les plus importants de ces technologies. Il ne faudra pas non plus nous interdire d'évoquer le futur de ce domaine et les actions que nous pourrions mener pour éviter certains excès et favoriser certaines évolutions.

1. Définir l'Intelligence Artificielle

Il peut sembler difficile de donner une définition unique pour qualifier un phénomène que l'on retrouve fréquemment dans des contextes très différents. On va parler de compagnon intelligent pour qualifier un robot qui va apprendre à vous servir mais on utilisera aussi l'adjectif intelligent pour qualifier une voiture capable d'évoluer de façon autonome ou même une machine à laver ou un appareil photographique capable de se régler seul. En dehors de notre monde tangible, on croise aussi souvent des logiciels intelligents sur internet, par exemple des programmes qui analysent nos habitudes de visite sur internet pour nous envoyer ensuite des propositions (voire des publicités) supposées être adaptées à nos besoins (comme les systèmes de recommandation).

En fait, tous ces systèmes (réels ou virtuels) peuvent être qualifiés d'agents intelligents dans la mesure où ils perçoivent leur environnement et y sélectionnent des actions pour atteindre un but. Cela suppose donc que ces agents soient capables de percevoir leur environnement (la route ou des pages web) et qu'ils soient donc équipés des capteurs adéquats. Cela suppose aussi qu'ils soient capables d'agir sur le monde, c'est à dire de le transformer, que ces actions soient réelles (freiner ou prendre un objet) ou virtuelles (décider d'envoyer un courrier électronique contenant une certaine publicité), de manière adaptée à la situation et non pas de manière automatique. Un moyen de s'assurer de ce caractère non stéréotypé est justement d'introduire un critère de succès, permettant de mesurer une performance par rapport à l'atteinte d'un but, que celui-ci soit réel (éviter un obstacle) ou virtuel (gagner la confiance et donc le financement des annonceurs). On peut remarquer à ce stade que cette définition, pour le moment assez générale, d'un agent intelligent n'est pas foncièrement éloignée de celle que l'on pourrait donner pour définir le comportement d'un être vivant que l'on qualifierait d'intelligent.

1.1 Pourquoi la question de l'Intelligence Artificielle est-elle difficile ?

On peut aussi remarquer, si l'on réfléchit maintenant à la possibilité de programmer un système numérique doté de capteurs, d'effecteurs et de critères de succès, que de façon très pragmatique, cette question pourrait se résumer ainsi : Est-on capable dans n'importe quelle situation, sur la seule base des informations reçues (c'est à dire pour chaque ensemble de signaux qu'il est possible de capter et en se donnant aussi la possibilité d'utiliser des états passés conservés en mémoire, des connaissances et des retours d'expérience associés) de choisir, dans le répertoire des actions possibles, la meilleure action à déclencher dans le sens où, même à un horizon lointain, elle apportera la meilleure contribution possible pour satisfaire le critère de succès ? Même si cette question peut paraître parfois assez critique (car il s'agit finalement de faire un programme permettant à l'agent de faire cette association Stimulus-Réponse sans l'aide d'un humain), sa formulation reste finalement assez simple et on peut donc se demander où exactement se cache la complexité de la réalisation d'agents

intelligents, autrement dit de savoir quels sont les sujets difficiles sur lesquels la recherche en Intelligence Artificielle travaille actuellement et sur lesquels elle bute depuis plus de soixante ans. On peut observer trois types de difficulté, que nous allons illustrer avec une étude de cas, où nous voudrions demander à notre robot-compagnon de nous apporter à boire.

1.1.1 Apprentissage et connaissances

Pour répondre à cette requète, le robot doit tout d'abord avoir accès à un certain nombre d'informations qui ne nous semblent évidentes qu'à cause de notre expérience de ce type de tâches. Il doit par exemple savoir que l'eau (mais pas le vinaigre, ni même l'eau contenue dans un vase contenant des fleurs coupées) peut satisfaire notre soif, qu'il peut être pertinent de trouver dans son environnement une bouteille d'eau minérale ou un robinet (mais dans ce dernier cas, qu'il doit aussi trouver un récipient pour transporter l'eau), etc. Tous ces exemples indiquent le besoin d'avoir recours à des valeurs et à des règles (par exemple de l'eau peut être bue, un verre peut avoir certaines formes et peut contenir de l'eau, etc.). Ces valeurs et ces règles peuvent être acquises par apprentissage ou fournies directement sous forme de connaissances a priori, ce qui va indiquer deux processus fondamentaux à élaborer (des capacités d'apprentissage et des capacités de représentation et de manipulation de connaissances) et leurs difficultés associées (comment assurer l'entrainement préalable adéquat ou comment s'assurer avoir fourni toutes les connaissances qui pourraient être utiles ?).

1.1.2 Résolution de problèmes

Même avec un exemple aussi simple que celui proposé ici, on peut voir qu'une résolution de problème induit souvent une décomposition temporelle et hiérarchique. Hiérarchique car fournir de l'eau peut se décomposer en trouver l'endroit où se trouve le robinet et l'endroit où se trouve le verre, chacun de ces sous-buts pouvant lui-même éventuellement être affiné (ouvrir la porte du placard pour prendre le verre). Temporelle car l'organisation dans le temps du comportement est ici essentielle (il est inutile d'aller vers le robinet tant qu'on n'a pas pris le verre ; quand on se dirige vers le verre, il ne faut pas oublier que le but principal est ensuite d'aller le remplir d'eau). Ici aussi, on peut constater un certain nombre de difficultés associées, pour explorer et organiser, dans l'espace et dans le temps, l'ensemble des actions qu'il faut réaliser pour atteindre son but, dont l'étude se retrouve dans ce que l'on appelle le domaine de la résolution de problèmes.

1.1.3 Heuristiques et induction

Il se trouve également que le nombre d'objets dans l'environnement et que le nombre d'actions possibles, ainsi que des valeurs et des règles qui leur sont associées peuvent être tellement grands qu'il ne sera pas possible de considérer tous les cas possibles dans la recherche de solution. Ainsi, votre robot-compagnon ne va pas pouvoir explorer chaque centimètre carré de son environnement qui peut être très vaste, pour y trouver une bouteille ou un robinet. Il peut se trouver aussi que ces observations soient entachées d'inexactitudes parce qu'on a mal perçu quelque chose ou parce que les règles que l'on exploitait auparavant ont changé : le monde est incertain. La perception de la bouteille peut être occultée par un autre objet ou il peut s'avérer qu'elle contient finalement du vinaigre. Quoiqu'il en soit, il est

important, pour des applications réalistes, qu'un agent intelligent soit capable d'évoluer dans un monde incertain et où la combinatoire empêche de considérer l'exhaustivité des cas. Ses méthodes de recherche de solution devront prendre en compte ces caractéristiques. Cela veut dire en particulier que ce seront des heuristiques, où le but est de trouver en un temps raisonnable une solution raisonnable plutôt que de garantir la solution optimale mais au prix d'une recherche dont la durée pourrait être prohibitive. Cela veut dire également que la raisonnement utilisé aura plutôt les caractéristiques de l'induction, très fréquemment utilisé par les humains aussi, où l'on va généraliser nos conclusions, à partir de l'observation de quelques cas particuliers.

Ce troisième type de difficultés relatif à la combinatoire et aux incertitudes est peut-être le moins évident à mettre en avant (et on trouve des études en Intelligence Artificielle qui incluent des recherches exhaustives et des processus de déduction) ; c'est pourtant celui qui va le mieux caractériser l'Intelligence Artificielle et différencier ce domaine d'autres domaines des mathématiques appliquées comme l'optimisation où on sera par contre plutôt intéressé par pouvoir prouver des propriétés d'optimalité ou par fournir des garanties de convergence: l'Humain est un satisfaiseur (il satisfait des contraintes), plutôt qu'un optimiseur, disait Herbert Simon, un des pères fondateurs de l'Intelligence Artificielle.

1.2 Les deux grandes approches de l'Intelligence Artificielle

Développer une intelligence artificielle, c'est à dire un système capable de percevoir un problème, à travers ses perceptions, sa connaissance et son expérience, et d'y apporter une réponse adaptée, par un comportement organisé dans le temps et l'espace, répondant à des critères de performances, est une question à laquelle de nombreux chercheurs essaient de répondre depuis au moins soixante ans. Pour ce faire, ils ont développé des travaux d'une grande variété, mais on peut tout de même essayer de les organiser autour de deux types d'approches représentatives.

1.2.1 Intelligence Artificielle symbolique

Une première approche que nous appellerons Intelligence Artificielle symbolique repose historiquement sur ce qui s'appelle l'hypothèse du système de symboles physiques. Il y est proposé que le monde peut être décrit par des formes physiques (des symboles), combinées en structures (des expressions) et manipulées (par des processus) pour produire de nouvelles expressions. Ce type d'approches stipule que la pensée humaine est une manipulation de symboles et remarque que l'ordinateur (qui manipule lui aussi des symboles) est particulièrement bien adapté pour réaliser de tels processus. De nombreuses réalisations ont été proposées dans cette direction et ont montré des résultats intéressants dans les domaines de la logique et de la résolution de problèmes. On peut citer en particulier les fameux systèmes experts ou encore des techniques de preuve de théorèmes, ainsi que les heuristiques associées aux parcours de graphes.

Les faiblesses de ces systèmes sont principalement relatives au fait que, si la puissance de raisonnement est présente, on suppose ici que les connaissances a priori nécessaires (que l'on appelle aussi parfois le modèle du monde) peuvent être spécifiées et fournies par l'utilisateur humain. Si ceci est réaliste dans le cas d'un système expert spécialisé, par exemple dans un

domaine médical précis, les connaissances dites de sens commun auxquelles on fait appel pour résoudre les problèmes dans la vie de tous les jours sont plus difficilement formalisables. Une autre limitation importante, liée à cette dernière, est directement issue de l'hypothèse du système de symboles physiques. Ancrer des symboles abstraits dans la réalité est souvent difficile et une description sub-symbolique s'accorde mal à la logique de cette approche, sans compter les domaines plutôt sujets à des traitements non symboliques comme par exemple les émotions ou la mémoire implicite.

1.2.2 Intelligence Artificielle numérique

De façon contrastée par rapport à cette approche logique et formelle de l'intelligence, l'approche numérique prend un point de vue nettement différent et se base avant tout sur les données du monde réel et sur les techniques d'inférence probabiliste plus à même de les traiter. Les approches principales dans ces domaines sont toutes les familles de réseaux de neurones, ainsi que les approches bayésiennes et statistiques. Leurs succès majeurs sont dûs aux algorithmes d'apprentissage automatique qui y sont associés et qui sont la raison majeure du renouveau de l'Intelligence Artificielle que nous vivons aujourd'hui.

Les limitations principales de ces approches numériques sont d'abord relatives aux capacités de raisonnement limitées de ces systèmes numériques, travaillant de plus sur des données représentées à plat, sans structure. Ensuite, elles ont des liens trop faibles avec les connaissances, niveau de représentation important pour communiquer avec l'expertise humaine. Il n'est ainsi pas toujours aisé d'introduire des connaissances a priori dans un système numérique, ce qui représente pourtant le seul moyen de ne pas confondre corrélation et causalité. Il est réciproquement tout aussi difficile d'extraire des connaissances exploitables de systèmes numériques comme les réseaux de neurones, ce qui serait pourtant un moyen de choix pour leur donner des capacités d'explicabilité, particulièrement importantes pour des applications critiques où l'utilisateur recherche des garanties avant de suivre la prescription donnée par un outil numérique prenant l'apparence d'une boite noire.

1.2.3 Positionnement

Ces deux approches, numériques et symboliques, ont ainsi actuellement chacune leurs faiblesses. On peut essayer de choisir préférentiellement des domaines applicatifs qui vont être bien adaptés à l'une ou à l'autre approche ; on peut aussi s'attacher à proposer des améliorations algorithmiques qui vont pousser un peu plus loin les performances et les domaines de compétence de chaque approche. C'est ce à quoi s'emploie principalement la recherche aujourd'hui.

Une autre stratégie consiste à essayer de tirer profit de ces deux approches et des propriétés qui leur sont associées. C'est ce qu'on peut faire par exemple en développant des systèmes dits hybrides, qui associent et font communiquer des modules des deux sortes. Une autre voie consiste à continuer à viser le développement de systèmes intelligents aussi bien capables de raisonnement et de manipulation de connaissances abstraites que d'apprentissage et de traitement de données à grain fin, en faisant en particulier remarquer que l'intelligence naturelle, animale et humaine, est un très bon exemple de système de traitement

d'information jouant sur tous ces tableaux. La référence à la cognition animale et humaine et parfois aux neurociences devient alors une autre voie d'approche de l'Intelligence Artificielle.

2 Le lien avec l'intelligence naturelle

En fait, le lien entre les intelligences naturelles et artificielles est un peu plus complexe que ça et de nombreux chercheurs font remarquer qu'ils prétendent faire de l'Intelligence Artificielle sans aucune référence à la cognition ni au cerveau. Ils pourront alors rencontrer quelques unes des limitations mentionnées plus haut mais ils pourront chercher à les compenser en insistant sur d'autres caractéristiques délibérément non naturelles comme des puissances de calcul ou de stockage disproportionnées. Ces approches peuvent ainsi rencontrer certains succès (comme c'est le cas actuellement avec le deep learning) mais on pourra alors se demander si on reste bien là dans le domaine de l'Intelligence Artificielle tel qu'il a été défini historiquement. Inversement, et on discutera ce point plus loin dans ce texte, le lien entre intelligence naturelle et artificielle peut devenir trop ténu et certains peuvent être tentés d'aller au delà de la référence et de la simple analogie pour ne plus les distinguer et finir par faire des généralisations et des amalgames qui peuvent se réléver dangereux ou au moins invalider la solidité de la démarche scientifique. Il conviendra donc de garder la bonne distance entre des algorithmes numériques que l'on cherche à implanter sur un système mécanique et électronique et la complexité du vivant, sans pour autant s'interdire de comparer les caractéristiques de traitement de l'information de ces deux types de systèmes.

2.1 Quelques points de comparaison

Lorsqu'on essaie de comparer les modes de traitement de l'information exploités par des systèmes naturels ou artificiels, différentes oppositions sautent rapidement aux yeux. Tout d'abord, il a été relevé depuis longtemps la distinction entre une intelligence dite géométrique, rigoureuse, analytique, excluant la contradiction et l'affectivité et largement utilisée dans le raisonnement logique, et une intelligence plus intuitive et plus globale, dite émotionnelle, qui nous permet parfois de prendre des décisions rapidement, en nous reposant sur des valeurs affectives que nous pouvons mettre sur certains choses et qui est un très bon moyen d'ancrer ces choses dans le réel.

Ensuite, nous avons pour le moment évoqué des capacités de raisonnement en faisant référence à la résolution de problèmes et plus généralement à des comportements dirigés par des buts du monde réel. Or, nous pouvons aussi constater que, le plus souvent, nos comportements sont dirigés par des motivations ou d'autres processus internes, ce qui nous donne notre autonomie et ne nous limite pas à une dépendance à notre environnement.

Enfin, des mécanismes comme l'intentionalité, l'imagination ou encore la créativité sont abordés dans les neurosciences cognitives, alors qu'ils sont singulièrement absents de l'Intelligence Artificielle. On ne peut pourtant pas nier qu'ils sont parties prenantes de nos compétences cognitives et qu'ils jouent très certainement un rôle dans la flexibilité de ces fonctions. Leur absence des attendus de l'Intelligence Artificielle souligne le manque d'exhaustivité de cette dernière.

Tous ces mécanismes, avérés dans la cognition, sont absents de l'Intelligence Artificielle et ils sont de plus liés à certaines de ses faiblesses mentionnées plus haut. En y regardant de plus près, on peut en fait se rendre compte que considérer ces mécanismes, c'est considérer le rôle du corps et du monde intérieur dans la cognition et, dans les neurosciences, leurs liens avec le cerveau. Vouloir ouvrir l'Intelligence Artificielle à ces propriétés essentielles de l'intelligence naturelle est donc une motivation majeure pour un travail bio-inspiré, visant à élucider les bases neuronales des fonctions cognitives en n'oubliant pas le corps, que cela soit à travers la place qu'il occupe dans l'espace ou les fonctions physiologiques qu'il assure.

2.2 Quelques principes cognitifs et neuronaux

Redynamiser la recherche en Intelligence Artificielle peut effectivement passer par revisiter quelques concepts issus des sciences cognitives, incluant aussi les neurosciences. Les liens entre données et connaissances, l'ancrage de ces dernières dans le monde réel et l'existence d'un sens commun ou d'un savoir faire implicite sont une première gamme de sujets à aborder pour renouveler l'Intelligence Artificielle. Ils peuvent trouver leurs bases entre les différentes formes de mémoires, implicites et explicites, qui existent dans notre cognition, et dans les mécanismes neuronaux qui permettent la communication et les échanges entre ces mémoires. Ils sont également étroitement associés à l'incarnation de l'intelligence naturelle, qui permet d'ancrer la connaissance au ressenti du corps et à ses besoins fondamentaux, matériels et spirituels. Comme évoqué plus haut, pouvoir utiliser les notions importantes d'émotions et de motivations dans la définition d'un comportement intelligent doit trouver ses racines dans la définition d'un corps ayant des besoins et ressentant plaisirs et douleurs.

Une autre gamme de sujets doit considérer les conditions de l'autonomie d'une intelligence qui se libère de sa dépendance à l'environnement pour définir ses propres buts mais aussi être capable de détecter les incertitudes associées. C'est en particulier l'action qui permet de nous mettre à la bonne distance de l'environnement, surtout si, dans un point de vue énactif, on la considère comme un moyen non pas de comprendre le monde mais d'interagir avec lui. Le comportement serait alors le contrôle de la perception par les actions et ceci permet de considérer non seulement la finalité de l'action (son but), mais aussi son intentionnalité, c'est à dire les moyens qu'elle utilisera. Par ailleurs, la pensée deviendrait ainsi un concept accessible, en la considérant, de façon similaire, comme le contrôle de l'imagination par des actions simulées.

S'intéresser à l'intelligence naturelle, c'est aussi s'intéresser à ses limitations. Pas nécessairement pour essayer de les compenser mais plutôt pour les mettre en contexte et essayer d'en comprendre les fondements. Il est ainsi parfois compréhensible de fournir une réponse de qualité moindre si elle peut s'expliquer par des compensations (en temps, en énergie ou selon d'autres dimensions) qui n'ont pas le même sens pour un système artificiel. Dans ce cadre, comprendre comment nous raisonnons en complexité limitée est un sujet important pour renouveler l'Intelligence Artificielle qui a trop souvent abordé ce thème sous l'angle de la logique alors que de nombreux indices nous montrent que nous ne sommes pas logiques dans nos raisonnements mais plutôt probabilistes. Suivre les principes mathématiques de l'inférence Bayésienne ou des processus de Dirichlet serait dans ce cadre une approche plus réaliste (on notera en particulier qu'ils font appel à la notion des connaissances a priori dont l'importance a déjà été soulignée plus haut) s'ils ne nécessitaient

des puissances de calcul et de stockage irréalistes dès que le problème devient de taille conséquente. Etudier comment approximer ces modèles mathématiques devient par contre un sujet majeur dans ce cadre.

3 Les enjeux de l'Intelligence Artificielle

Au delà de la description factuelle de l'état de l'Intelligence Artificielle, de ses faiblesses et de ses pistes d'évolution, que nous avons esquissée, il est aussi (et même probablement plus) important, considérant sa nature, de se questionner sur son impact sur notre société et sur ses évolutions possibles et d'évoquer des réflexions qui nous semblent pertinentes pour aborder les aspects éthiques associés.

3.1 Un premier bilan sur l'Intelligence Artificielle faible

Dans ce texte, nous avons d'une part évoqué l'Intelligence Artificielle telle qu'on la rencontre aujourd'hui, qu'on appelle aussi parfois l'Intelligence Artificielle faible parce qu'elle est spécialisée et se limite à certaines tâches (jouer aux échecs, aider à conduire une voiture, traduire un texte, etc.). Elle se présente comme nous permettant une vie plus facile mais elle pourrait aussi la rendre plus stéréotypée, plus uniforme. Ici un enjeu important est celui de la donnée qui doit à la fois être protégée quand elle a un caractère personnel et accessible au plus grand nombre (et non pas propriété exclusive des grandes entreprises du numérique) quand elle est un bien commun (des atlas routiers aux modèles de langage). Les lois européennes récentes (comme le RGPD) vont plutôt dans la bonne direction, ainsi que les souhaits de transparence des algorithmes et d'interprétabilité de certains processus artificiels, même si on peut se demander si ces défis seront toujours tenables.

Un autre point sur lequel il convient d'insister concernant l'Intelligence Artificielle faible concerne la manière de se présenter qui doit rester « humble », c'est à dire limitée factuellement aux compétences démontrée. Il ne faudra ainsi pas présenter un nouveau système d'aide à la marche capable de s'adapter dans certaines circonstances comme le système révolutionnaire qui va permettre à tout paraplégique de marcher sans effort et ainsi donner de faux espoirs et obscurcir l'image de ces nouvelles technologies.

Un autre moyen de limiter les fantasmes dans ce domaine est de se souvenir que les rapports de l'humain à la machine et à l'outil ne sont pas nouveaux. Faire des outils pour se dépasser est dans la nature humaine et même si nous avons aujourd'hui des machines qui se déplacent, cassent des cailloux ou font des calculs plus vite et mieux que nous, nous ne pensons pas pour autant qu'elles nous dépassent. En fait, nous ne comparons même pas tant il est évident que les machines nous sont complémentaires, qu'elles peuvent être améliorées sans remettre en cause notre intégrité et qu'elles font en fait partie de notre culture. Il en est de même pour les questions de responsabilité en cas d'accident impliquant une machine. Il ne viendrait pas à l'idée de lui donner une personnalité juridique et on cherchera plutôt des responsabilités du coté des humains qui l'ont construite, vendue, louée ou utilisée. On devrait raisonnablement pouvoir étendre cette manière de voir même si la machine a acquis une parcelle d'intelligence.

3.2 Que penser de l'Intelligence Artificielle forte ?

Dans ce texte nous avons d'autre part évoqué l'Intelligence Artificielle dite forte, qui n'existe pas aujourd'hui, en aucune manière, mais qui est celle à laquelle on fait référence quand nous parlons de regrouper l'ensemble de nos compétences cognitives dans un même système qui pourrait ainsi nous égaler voire nous surpasser. Même si ce thème est souvent abordé quand il s'agit de mieux nous comprendre (par exemple pour mieux nous soigner), la frontière n'est pas loin pour imaginer des machines qui pourraient nous remplacer et il est donc légitime de se demander si c'est là vers où on veut aller. On pourrait aussi se demander si c'est raisonnable de jouer ainsi à se faire peur et, finalement, si on pourra tout simplement atteindre un tel niveau de développement. Mais il semble tout de même utile d'aborder ce sujet, ne serait-ce qu'en invoquant ce que l'on appelle la loi d'Amara, qui énonce que, si on surestime généralement l'effet des technologies à court terme, on le sous-estime plutôt à long terme.

Quand on nous parle des dérives potentielles d'une Intelligence Artificielle forte (qui reste à venir), un premier point de vue est à évacuer rapidement car il relève du fantasme pur, c'est le transhumanisme et sa notion clé de singularité, ce moment où l'humain serait dépassé par la machine et ses conséquences potentielles (qui vont de l'éradication pure et simple de l'humanité à la naissance d'une post-humanité mécanique). C'est un fantasme pur car, quand on regarde les écrits prophétisant ces futurs, ils correspondent à une vision non-scientifique, ne reposant sur aucune méthode de travail scientifique, dualiste et sans référence à la notion de corps. La raison pour laquelle ce fantasme marque autant c'est qu'il est énoncé sous la forme d'une fatalité ; autrement dit, on ne nous dit pas « attention nous pourrions avoir ça », on nous dit « on va avoir ça ; comment s'y préparer au mieux en en tirant le plus de bénéfices possibles, par exemple en en profitant pour acquérir l'immortalité ». On peut éventuellement comprendre la base de cette argumentation quand on regarde qui la profére. Il s'agit parfois de chercheurs en quête de reconnaissance et de financement ou de grandes entreprises du numérique qui cherchent à se rendre ainsi bienveillantes et à masquer un autre enjeu, bien réel, que nous avons déjà évoqué, l'accès à nos données.

Il y a une autre dérive potentielle qui est bien plus sérieuse, et que nous connaissons déjà car elle est née avec notre société technologique, c'est le culte de l'efficacité tel qu'il a été théorisé par Jacques Ellul. Nous avons trop souvent une relation technique au monde, que nous voulons dominer, maitriser, optimiser et nous faisons de cette démarche un principe général, comme nous l'avons déjà démontré avec l'environnement. Replacé dans le cadre d'une telle volonté d'asservir le monde, l'Intelligence Artificielle peut devenir un outil redoutable s'il est utilisé aveuglément, juste car c'était possible et non car ça répondait à une demande particulière et réfléchie à l'avance. C'est probablement cette crainte qu'expriment certaines réticences au déploiment systématique et mal justifié de certains outils, qui se matérialisent en particulier selon deux critiques. La première est l'uniformisation et la deuxième est l'absence de réflexion éthique sur l'impact de ces dispositifs dans notre société. Concernant l'uniformisation, notre cerveau ayant déjà tendance à filtrer le bruit, des systèmes d'Intelligence Artificielle et leur approche statistique des données vont encore amplifier ce biais et nous faire voir le monde selon ses tendances principales et en oubliant sa diversité, ce que d'autres domaines scientifiques comme la biologie et les sciences sociales nous rapportent pourtant comme étant important de leurs points de vue. Quant à la réflexion éthique, elle est nécessaire pour prévenir une utilisation systématique d'une technologie pouvant avoir un impact aussi fort sur notre monde mais aussi sur notre cerveau lui-même et

elle devra inciter à se poser la question d'expliciter les bénéfices visés et de se demander s'ils sont effectivement compatibles aux valeurs que notre société aura définies au préalable.

3.3 Quelles pistes pour le futur ?

Comme beaucoup de technologies à fort impact, l'Intelligence Artificielle peut être une richesse pour la société si nous savons nous en servir, mais elle peut aussi augmenter ses inégalités et ses biais. C'est en ayant à l'esprit ce constat que l'on peut proposer quelques pistes pour le futur. La première recommandation est relative à la formation et à l'éducation. Chacun pourra mieux comprendre le potentiel et les risques de l'Intelligence Artificielle s'il sait de quoi il parle. La seconde recommandation concerne une ouverture multidisciplinaire. Tout d'abord, nous l'avons mentionné au début de ce texte, il y a une certaine proximité entre l'Intelligence Artificielle et les neurosciences et travailler dans ce dernier domaine peut se concevoir comme une voie pour mieux comprendre nos fonctions cognitives et donc permettre de contribuer au premier. En retour, son coté opératoire peut permettre de manipuler et de décortiquer ces fonctions, pour mieux comprendre le cerveau mais aussi ses dysfonctionnements. Ensuite, nous venons d'évoquer les sciences humaines et sociales et les sciences biologiques pour s'interroger sur le monde réel dans lequel on vit et sur les valeurs de la société dans laquelle nous voulons vivre. Très probablement, l'apport d'experts d'autres disciplines pourra nous être utile dans ces perspectives. La troisième recommandation insistera encore sur la nécessité d'associer une réflexion éthique au développement de la recherche en Intelligence Artificielle. Alors que j'ai alerté plus haut sur le risque d'amplifier les excès de la relation technique de l'humain au monde en développant aveuglément l'Intelligence Artificielle, inversement on peut aussi penser que, justement car ce qui est en jeu ici c'est notre cognition et son rapport à notre corps et au monde qui l'environne, cette réflexion éthique sera aussi l'occasion de nous demander collectivement ce qu'être humain veut dire.