Luc Steels

Les machines peuvent-elles avoir un libre arbitre ?

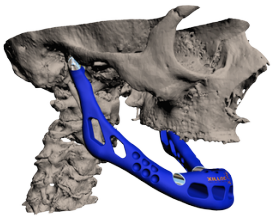
traduit de l'anglais par Valentine Vasak

Il y a 20 ans, j'ai écrit un article intitulé « L'Homo Cyber Sapiens et le Robot Homonidus Intelligens : la vie artificielle et son approche de l'intelligence artificielle »[1]. J'ai élaboré l'hypothèse selon laquelle le rapprochement entre l'intelligence humaine et celle des robots pourrait entraîner l'émergence de deux nouvelles espèces. La première, que j'ai nommée Homo Cyber Sapiens est destinée à succéder à notre propre espèce, l'Homo Sapiens. Elle se caractérisera par des extensions de notre corps et de notre cerveau : une mémoire améliorée par des procédés numériques, un traitement de l'information plus performant grâce à des logiciels d'intelligence artificielle, des modules permettant une connexion Internet directe en wifi, des capacités sensori-motrices accrues, comme par exemple un œil ou une main supplémentaires, ou encore l'utilisation d'exosquelettes robotiques pour optimiser les mouvements du corps. La seconde espèce, que j'ai appelée Robot Homonidus Intelligens renvoie à une génération future de robots humanoïdes avec un niveau d'intelligence comparable à celui des humains. Ces robots auraient une forme humanoïde et leurs capacités seraient presque équivalentes à celles des êtres humains pour ce qui est de la perception visuelle, du contrôle moteur, de la planification, de la représentation des connaissances, de la mémoire épisodique, et du traitement automatique du langage.

Au cours des vingt dernières années, la technologie n'a cessé de progresser, à tel point que ces réalités ne nous semblent plus si inaccessibles, et le rythme des avancées semble encore s'accélérer. D'un autre côté, ces spéculations restent clairement du domaine de la science-fiction. A bien des égards, nous sommes encore très loin de voir ces nouvelles espèces évoluer parmi nous. C'est en partie dû à nos technologies limitées et au fait que nous connaissons encore mal la nature de l'intelligence, mais cela s'explique également par le fait que ces développements soulèvent des questions de société cruciales qui doivent êtres posées en amont. En effet, nous devons à tout prix éviter de nous laisser déborder par ces avancées techniques, comme cela a si souvent été le cas par le passé.

Quels progrès technologiques avons-nous réalisés et quels domaines posent encore problème ?

1. Notre capacité à fabriquer des objets physiques s'est développée de manière fulgurante. Les imprimantes 3D constituent l'une des dernières avancées en la matière et permettent d' « imprimer » n'importe quel objet conçu par ordinateur en superposant successivement des couches de métal ou de plastique très fines et en chauffant des particules de poudre de métal grâce à des lasers de haute précision. Ce procédé est un dérivé des technologies d'impression laser et de fabrication de puces électroniques. L'impression 3D est devenue un outil de base abordable pour la fabrication de composants robotiques, et l'application de ces techniques permet d'importantes réparations des différentes composantes du corps humain. Le document 1 montre l'exemple d'une mâchoire inférieure imprimée sur mesure puis implantée sur un patient qui n'était plus capable de parler ou manger en raison d'une ostéomyélite (infection osseuse).



Document 1. L'impression 3D permet de fabriquer sur mesure avec une très grande précision certains éléments corporels et donc de réparer ou améliorer les performances du corps humain.

1. "The Homo Cyber Sapiens, the Robot Homonidus Intelligens and the Artificial Life approach to Artificial Intelligence.", Steels, 1994

2. De plus en plus souvent, les objets manufacturés sont équipés d'une mémoire et d'une puissance de traitement intégrées, ainsi que de capteurs et de mécanismes déclencheurs. Ces progrès ont été rendus possibles par une capacité de miniaturisation de plus en plus poussée et par la baisse du coût des technologies de traitement de l'information (telles que les processeurs ou les puces mémoires). Les avancées récentes qui permettent d' « imprimer » des circuits, des capteurs ou des déclencheurs grâce au procédé de l'impression 3D ont aussi joué un rôle crucial. En termes de conception et de production, l'impact de ces technologies est énorme. Dans le cas des robots, les données informatiques peuvent désormais être traitées séparément par plusieurs parties du corps à la fois, ce qui rend en principe les machines plus robustes. Le document 2 montre l'exemple d'un robot humanoïde construit récemment qui présente ces caractéristiques (Hild, et al. 2012). La jambe est capable de se lever toute seule, le bras peut bouger même s'il est détaché du buste, et le corps tout entier conserve son équilibre même si on dévisse la tête. L'intégration de technologies informatiques implantées directement dans les composantes manufacturées ouvre un champ d'opportunités incroyables en vue de l'augmentation des performances du corps humain. Cela vaut d'autant plus si l'on parvient à établir un lien direct entre les circuits neuronaux du système nerveux et les dispositifs artificiels qui permettent ces améliorations. On a pu observer récemment plusieurs démonstrations convaincantes de ce principe permettant au cerveau de recevoir des informations directement d'un implant cochléaire ou d'une rétine artificielle, ou encore une stimulation neurale directe par des mains artificielles.

Document 2. L'alimentation énergétique et le traitement des informations du robot MYON sont entièrement distribués, les différentes composantes du corps du robot peuvent donc fonctionner de manière autonome et il commence à avoir un comportement général cohérent.

3. Les logiciels-artefacts que nous sommes capables de fabriquer sont de plus en plus complexes, même si ces évolutions demeurent pour l'ensemble invisibles pour les utilisateurs en fin de chaîne de production. Désormais, on trouve vissés à l'oreille d'une grande partie de la population mondiale des téléphones portables bien plus performants du point de vue informatique et dotés d'une mémoire bien supérieure aux ordinateurs avec lesquels travaillaient les chercheurs à la pointe de la technologie dans les années soixante-dix. Ces portables incluent désormais des logiciels de traitement de l'image, de télécommunications et d'intelligence artificielle d'une complexité incroyable. Cette tendance constante à la complexification des logiciels s'explique par l'accumulation progressive de composants informatiques produits par une armée de développeurs qui sont réutilisés et réintégrés dans des systèmes à échelle de plus en plus grande. Ici encore, les applications dans la construction de robots autonomes ou de prothèses sont très nombreuses. Par exemple, certaines prothèses auditives offrent aujourd'hui un traitement du signal sonore et une reconnaissance des structures du langage extrêmement sophistiqués, de la même manière, les appareils de vision robotiques sont capables de traiter des informations visuelles très complexes en temps réel, et d'élaborer la cartographie d'un environnement dynamique même lorsqu'ils évoluent à l'intérieur de cet environnement.

4. Le traitement des informations à l'échelle locale est également de plus en plus souvent associé à des ressources en ligne accessibles grâce au principe du « cloud computing » (ou informatique en nuage) qui met à la disposition de l'utilisateur une mémoire de stockage importante et une puissance informatique conséquente même pour des outils très simples. Des applications telles que SIRI sur l'Iphone d'Apple constituent déjà un exemple de la manière

dont cette technologie peut conduire à une intelligence augmentée. C'est l'une des modalités permettant aux robots d'accroître comme jamais auparavant leur capacité de traitement des informations. Si l'on parvenait à mettre en relation le cerveau humain et ces vastes ressources informatiques, nos capacités mentales pourraient être améliorées de manière considérable.

Néanmoins, malgré toutes ces avancées technologiques, des obstacles techniques d'envergure rendent impossibles les changements radicaux qui pourraient entraîner une transition vers l'Homo Cyber Sapiens ou le Robot Homonidus Intelligens décrits par les auteurs de science fiction. Je crois tout d'abord qu'il est nécessaire d'évoluer d'une conception technique des systèmes à une conception biologique. La conception technique est basée sur le contrôle hiérarchique, une modularité stricte et une approche « top-down » (ou technologie descendante). La conception biologique est basée sur un contrôle distribué, des interactions non-modulaires, l'auto-organisation et l'évolution. L'idée est de « cultiver » des systèmes extrêmement complexes plutôt que de les concevoir. Le champ scientifique de l'intelligence artificielle s'est développé vers la fin des années 1990 : il s'agit d'explorer cette voie biologique afin de concevoir et de mettre en place des artefacts complexes (Steels and Brooks, 1994), mais la discipline peine à se généraliser et il existe relativement peu de programmes de recherche sérieux qui explorent ce domaine en profondeur. Deuxièmement, je pense qu'il est nécessaire d'investir bien davantage dans la recherche sur l'intelligence artificielle. Elle a en effet connu un formidable essor dans les années 1970 et 1980 et est à l'origine des technologies qui constituent aujourd'hui les fondamentaux d'Internet, en particulier le traitement automatique du langage naturel, et les technologies de recherches. Cependant, durant les vingt dernières années, les crédits alloués à la recherche fondamentale sur l'intelligence artificielle se sont taris, notamment en ce qui concerne les programmes de recherches les plus importants qui sont désormais financés par l'Union Européenne. Pourtant, nous sommes encore bien loin de comprendre parfaitement les fondements de la discipline. Nos modèles informatiques de traitement et d'apprentissage du langage restent très superficiels comparés à l'intelligence basée sur des structures sémantiques profondes qui caractérise notre espèce.

En dehors de ces limites techniques, ces technologies soulèvent des questions de société cruciales qui doivent être débattues. Pour atteindre la complexité requise au développement de formes très poussées d'intelligence artificielle, nous devons en tant qu'êtres humains céder une autonomie de plus en plus importante à des systèmes artificiels. D'une part, parce que les décisions prises sont si rapides qu'il est impossible de consulter un être humain pour qu'il examine les différents éléments en présence et entérine les décisions avant leur application. D'autre part parce que le processus de prise de décision est si complexe que son issue n'est plus entièrement prévisible. Ces avancées soulèvent des questions fondamentales en termes de responsabilité et d'imputabilité. L' Homo Cyber Sapiens doit-il être tenu responsable et puni tout comme nous, et ce même si une part importante des décisions prises et des actions menées le sont par des mécanismes autonomes qui optimisent son potentiel comportemental ? Un Robot Homonidus Intelligens peut-il être tenu responsable et puni bien qu'il ne soit au fond qu'une machine ? Il est urgent de répondre à ces questions. Pouvons-nous laisser une voiture autonome déclarée « prête pour la route » par les projets de recherches menés à Berlin ou Stanford conduire librement dans nos rues ? Que faire si ce véhicule est impliqué dans un accident ? Actuellement, ces questions freinent la mise en circulation des voitures-robots, alors qu'aujourd'hui 40 000 personnes par an sont tuées sur les routes européennes : peut-être que finalement, les machines se révèleraient des conducteurs plus fiables, même si elles ne sont pas infaillibles ?

Le cadre juridique en vigueur prévoit trois modalités punitives pour ce type de comportements. Premièrement, quelqu'un (un fabricant par exemple) peut être tenu responsable si un système défectueux entraîne des conséquences néfastes. Bien évidemment, les défaillances techniques doivent êtres prouvées et il peut y avoir des circonstances atténuantes. Dans le cas des systèmes autonomes, la défaillance est plus difficilement imputable au fabricant. Par exemple, prenons le cas d'une machine capable d'intégrer des procédés par apprentissage à laquelle le propriétaire doit fournir une partie des données nécessaires à cet apprentissage : si ces données sont insuffisantes et ne permettent pas de traiter tous les cas qui doivent être pris en compte, la question de la responsabilité se pose. La seconde modalité punitive met en cause le propriétaire du système si une mauvaise utilisation de la machine entraîne un préjudice. Une fois encore, en raison de cette augmentation de l'autonomie du système, il est difficile de désigner aisément un responsable de cette mauvaise utilisation. Troisièmement et il s'agit sans doute du cas le plus intéressant, une personne est punie parce qu'elle porte préjudice VOLONTAIREMENT, ce

qui implique que cette personne est douée d'un libre arbitre. Il ou elle agit intentionnellement dans un but précis et tout en sachant que son action aura des conséquences néfastes. Ces agissements rentrent dans la catégorie des actes criminels et sont punis en conséquence.

Peut-on considérer qu'un robot autonome et intelligent dispose d'un libre arbitre et puisse nuire intentionnellement ? La question est complexe. Le philosophe Dennett soutient que l'intentionnalité, les convictions, le potentiel d'action, la responsabilité et même la conscience sont des propriétés que nous attribuons aux autres agents pour nous permettre de comprendre et prédire leurs comportements. En d'autres termes, ce ne sont pas des propriétés des agents eux-mêmes, mais des propriétés de notre relation avec ces agents, et par conséquent, aucune raison intrinsèque n'empêche théoriquement l'agent d'être une machine. C'est ce qu'on appelle « la stratégie de l'interprète » (« the intentional stance » en anglais, Dennett, 1996). Mais pour de nombreuses personnes, la question n'est en fait pas si simple. Aujourd'hui déjà, nous attribuons toutes sortes de propriétés à nos machines, par exemple, quand nous disons que notre ordinateur ne veut pas s'allumer. Mais il s'agit ici d'une façon de parler, de là à être convaincu que la machine a réellement des convictions, des désirs et des intentions, il y a un fossé que tous ne sont pas prêts à franchir.

Quelque que soit la situation, nous devons commencer à nous demander si nous devons ou non faire des robots des personnes légales, ce qui impliquerait potentiellement qu'ils formeraient une nouvelles catégorie d'êtres dotés de droits constitutionnels. Des débats juridiques seront alors nécessaires, qui rappelleront peut-être ceux qui ont eu lieu dans la Rome Antique pour déterminer les droits, et la responsabilité juridique et morale des esclaves (Pagallo, 2013). Nous devons commencer à statuer sur la responsabilité juridique des robots : leur implanter des composants qui produiraient un raisonnement moral intégrant les limites imposées par les lois et les conventions en vigueur pourrait être une solution. C'est précisément ce que Arkin (2009) préconise à l'égard des robots militaires. Enfin, nous devons nous demander si les êtres humains doivent endosser de nouvelles responsabilités concernant le comportement des machines.

Références:

1. Arkin, R. (2009) "Governing lethal behavior in autonomous robots." Taylor and Francis, New York.

2. Dennett, D. (1996) *The Intentional Stance*. Cambridge, Massachusetts: The MIT Press.

3. Hild, M., T. Siedel, C. Beckendorff, C. Thiele, and M. Spranger (2012) "Myon, a new humanoid." In Steels, L. and M. Hild (eds.) (2012) *Language Grounding in Robots*. Springer Verlag, Berlin.

4. Pagallo, U. (2013) *The laws of robots. Crimes, contracts and torts*. Springer Verlag, Berlin.

5. Steels, L. (1994) "The Homo Cyber Sapiens the Robot HomonidusIntelligens and the Artificial Life approach to Artificial Intelligence." (publié en allemand sous le titre Steels, L. (1995) "Homo cyber-sapiens oderRobohominidusintelligens: MaschinenerwachenzukunstlichemLeben", Maar, C., E. Poppel and T. Christaller (eds) *Die Technik auf demWegzurSeele: Proceedings of theBurda Symposium on Brain-Computer Interfaces*. RowohltTaschenbuchVerslag. Hamburg. pp 327-344.)

6. Steels, L. and R. Brooks (1994) "The Artificial Life route to Artificial Intelligence: Building Situated Embodied Agents." Lawrence Erlbaum Ass. New Haven

---

Valentine Vasak prépare actuellement un doctorat sur l'œuvre du dramaturge américain Edward Albee à l'université de Paris IV (Sorbonne). Agrégée d'anglais, elle réalise régulièrement des traductions dans plusieurs domaines (sous-titrages de films, articles...).