Giornale di filosofia, vol. 2/2021 • ISSN 1827-5834 • DOI 10.7413/1827-5834012

Gilles Lecerf [*]

Intelligence artificielle, délégation de tâches et dégradation des facultés humaines: une réactualisation de la "honte prométhéenne"?

> L'I.A. (...) doit être en mesure de communiquer rapidement.
> Et notre rythme naturel de communication est très lent,
> surtout notre débit de sortie.
> (L'objectif est d')améliorer la bande passante de notre communication.
>
> Elon Musk, à propos de son entreprise Neuralink

Dans cet extrait tiré de son interview avec la journaliste Kara Swisher[1], Elon Musk évoque son inquiétude d'être dépassé par les créations technologiques contemporaines, en particulier les services d'intelligence artificielle (IA). Ce sentiment de relégation qui le fait se sentir tantôt semblable à une fourmi qu'on écrase sans y prêter attention ou à un arbre relativement mutique, s'apparente à une forme de honte. Honte d'une nature physique limitée, d'un corps organique qui dépérit, d'aptitudes bornées. Face à la machine qui permet de traiter l'information à une vitesse sur-humaine, Musk éprouve cette honte et propose une piste pour tenter d'en sortir: s'aligner avec la capacité de la machine à travers Neuralink, l'une des nombreuses entreprises dirigées par Musk, et dont l'objectif est d'interfacer le cerveau avec l'ordinateur pour les faire entrer en communication directe.

> Il n'est pas question pour l'homme contemporain de se résigner une fois pour toute à son infériorité et à son retard en acceptant le caractère borné de son corps. Il doit donc faire quelque chose. Son rêve serait évidemment de devenir semblable à ses dieux, les machines, ou, mieux encore, de leur appartenir au point de leur devenir en quelque sorte totalement et absolument consubstantiel[2].

[*] Doctorant en philosophie. Paris 1 Panthéon Sorbonne

[1] K. Swisher, *Opinion | Elon Musk: 'A.I. Doesn't Need to Hate Us to Destroy Us*, in "The New York Times", rubrique "Opinion", 28 septembre 2020 (en ligne: https://www.nytimes.com/2020/09/28/opinion/sway-kara-swisher-elon-musk.html).

[2] G. Anders, *L'obsolescence de l'homme: sur l'âme à l'époque de la deuxième révolution industrielle*, Éd. de l'Encyclopédie des nuisances, Paris 1956, p. 53.

Dans ce recueil au titre éloquent et publié en 1956, le philosophe Günter Anders définissait un concept qui nous semble aujourd'hui éclairant pour penser le mouvement technologique contemporain: la honte prométhéenne. Loin de la figure libératrice et mythologique du titan, condamné pour l'éternité pour avoir dérobé le feu et ainsi offert aux humains la possibilité de s'extirper de leur condition fragile et précaire, Anders souhaitait définir à nouveau le rapport entre l'homme et ses créations. Car après l'horreur des camps et des bombardements nucléaires, l'heure n'était plus à l'optimisme et à la naïveté.

Ce que nous allons tenter de faire dans ce travail, c'est nous interroger sur la pertinence de ce concept au sein du mouvement technologique contemporain.

Nous allons en particulier revenir sur les deux grandes thèses qui se rattachent à la honte prométhéenne:

– "nous ne sommes pas de taille à nous mesurer à la perfection de nos produits"[3];

– "ce que nous produisons excède notre capacité de représentation et notre responsabilité"[4];

Pour chacune d'entre elles nous allons essayer de montrer les phénomènes propres au mouvement technologique et qui permettent de penser la réactualisation de ce concept. Nous nous attacherons en particulier au processus de délégation des tâches et de dégradation des compétences humaines qui s'opère dans ce mouvement technologique. Enfin nous travaillerons sur les limites de cette représentation surhumaine des services d'intelligence artificielle afin de discuter de la possibilité de ménager une place à la singularité humaine dans ce mouvement technologique.

Honte prométhéenne et alignement

Lorsqu'Anders définit cette honte prométhéenne, il veut souligner qu'elle ne se limite pas à une simple honte devant les conséquences néfastes et directes de certaines productions, comme cela peut être le cas avec une arme. Pour Anders, la honte prométhéenne est fondamentalement une honte de nous-mêmes, de notre propre condition, devant, non pas la dangerosité, mais l'efficacité, l'optimisation et la performance de nos outils. Autant d'éléments objectifs, quantifiables, qui nous renvoient à notre propre défaillance, à notre propre faillibilité. Ce dont nous avons honte, pour Anders, c'est de notre condition fragile, de notre incapacité

[3] Ivi, p. 11.

[4] G. Anders, *L'obsolescence de l'homme: sur l'âme à l'époque de la deuxième révolution industrielle*, Éd. de l'Encyclopédie des nuisances, Paris 1956, p. 11.

à nous mesurer à l'efficacité, à la performance des machines que nous créons. Anders rappelle bien que l'image de Prométhée était pourtant celle d'une force libératrice illimitée, celle de "produire toujours du nouveau"[5]. Néanmoins il estime que notre absence de pensée critique sur cette liberté prométhéenne a progressivement créé un fossé béant entre notre capacité de production et notre capacité à "se figurer les conséquences de ce que nous avons nous-mêmes fabriqué"[6]. C'est cette "a-synchronicité chaque jour croissante entre l'homme et le monde qu'il produit (…)"[7] qu'Anders va appeler le décalage prométhéen.

Avec l'essor actuel de services d'IA dans de nombreux domaines, de la médecine, aux transports, en passant par la finance, le recrutement mais également la production de texte et les jeux de stratégie, la pertinence de ce concept de honte prométhéenne reste vivace. D'une part car la puissance de traitement des machines a été démultipliée depuis l'époque d'Anders et ce décalage entre nos capacités humaines et celles des machines n'a cessé de se creuser. D'autre part car ces services se propagent dorénavant à des domaines qui ne sont pas conditionnés à une force brute de calculs. Ainsi, si Kasparov a pu être battu par le logiciel DeepBlue d'IBM en 1996, c'est uniquement parce que la modélisation de l'ensemble des combinaisons possibles était possible aux échecs. La victoire d'Alphago, le programme de DeepMind face au champion de go Lee Sedol marque elle un tournant, celui de la simulation d'une forme de créativité dans un jeu qui ne peut être entièrement réduit à un calcul de probabilité.

Si Anders estimait que la performance surhumaine des machines pouvait causer cette honte prométhéenne, il est donc éclairant de s'interroger sur cette notion à l'heure où ces machines investissent des domaines que l'on pensait alors uniquement réservés aux humains et les y surpassent.

Discussion autour des réfutations soumises à Anders

Intéressons-nous aux critiques auxquelles Anders a répondu et à leur actuelle validité. La première critique est avant tout celle de la fierté prométhéenne. Non, l'homme moderne ne ressentirait pas de honte mais serait bien plutôt toujours empreint de cette fierté liée à sa capacité de production, d'invention, de maîtrise technologique. Si Anders ne nie pas que cette fierté puisse exister, il estime néanmoins que celle-ci est très limitée et se cantonne aux quelques experts qui ont justement produit la

[5] Ivi, p. 30.
[6] Ivi, p. 32.
[7] *Ibid.*

machine en question. "Le monde des instruments n'appartient pas (…) à ceux qui ne sont pas intégrés dans le processus de production"[8]. A l'heure actuelle, nous retrouvons ce même type de discours parmi les ingénieurs victorieux de l'équipe de DeepMind qui, face à un Lee Sedol défait, ressentent la fierté d'avoir conçu un tel programme doté de "performance surhumaine"[9]. Néanmoins cette fierté reste cantonnée aux rares privilégiés qui ont accès à cette technologie. Ce que l'on peut néanmoins dire sur cette fierté est qu'elle se mue, pour Anders, en une forme d'orgueil qu'il rapproche de l'*hybris. Hybris* qui devient l'un des moteurs principaux de cette honte prométhéenne. Ces experts fiers de leur liberté prométhéenne et animés par cette capacité infinie qui leur est offerte sont en réalité insouciants face au décalage qui se crée entre leurs créations et les humains. Pour Anders, la figure de cet expert est en réalité l'incarnation de "la présomptueuse auto-humiliation, (la) soumission animée par une volonté d'hybris"[10].

Si Anders refuse la possibilité d'une fierté partagée, la seconde critique qu'on lui adresse reste néanmoins que, à défaut de cette fierté, il n'existe pas non plus de honte. Que cette honte ne se manifeste jamais et qu'elle est donc fantasmée. Anders ne refuse pas non plus cette objection de l'invisibilité de la honte mais elle est simplement, pour lui, le signe de la dissimulation: "au lieu de chercher à dissimuler son opprobre et à disparaître, c'est désormais sa propre honte qu'il dissimule"[11]. Aujourd'hui, il semblerait que cette absence concrète de sentiment de honte puisse s'expliquer par deux éléments: d'une part car nous ne sommes pas personnellement confrontés à ce type de rencontre honteuse qui pourrait révéler notre faiblesse et balayer, en quelques lignes de code, l'apprentissage et les efforts consentis parfois sur toute une vie. De même que nous ne pouvons ressentir la fierté des ingénieurs, nous ne pourrions que difficilement ressentir la honte qu'éprouve Lee Sedol lorsqu'il doit admettre sa large défaite. Néanmoins il est clair que Lee Sedol l'a ressenti et son récent départ à la retraite, alors qu'il estimait qu'il n'était maintenant plus possible de pouvoir prétendre à se confronter à la machine, démontre que cette blessure honteuse peut être un véritable frein à l'action humaine. D'autre part car elle se dissimule également sous la forme d'une certaine fascination devant l'émergence de cette puissance surhumaine. Les spectateurs divertis seraient eux tout bonnement fascinés par ce qui

[8] G. Anders, *op. cit.*, p. 43.
[9] D. Silver, *et al.*, *Mastering the game of Go without human knowledge*, in "Nature", 550, 7676, octobre 2017, pp. 354-359.
[10] G. Anders, *op. cit.*, p. 67.
[11] Ivi, p. 44.

est en train d'advenir, dopés par un imaginaire cybernétique où la fascination de la machine intelligente est omniprésente.

La troisième objection qu'il discute est celle qui consiste à dire que les individus n'auraient pas honte de leurs créations mais bien plutôt du fait que, dans le monde industriel qui était celui d'Anders, ils se sentent réifiés par des processus productifs. Ils se sentiraient justement honteux d'être réduits à une machine basique, réduits à un simple élément d'une chaine de production cadencée et optimisée. Les individus se sentiraient rabaissés à cette machine productive, et non considérés pour leur singularité humaine. La honte serait donc justement un sentiment lié à l'instinct de conservation et de défense de la condition humaine face à la réification, et non un sentiment d'infériorité face à une machine. Mais pour Anders, ces deux moments de la honte ne sont pas contradictoires, ils sont simplement consécutifs. La honte prométhéenne qu'il décrit serait la phase suivante et consiste en ce "deuxième degré de l'histoire de la réification"[12] de l'homme. Après avoir effectivement ressenti une honte face à un processus qui cherche à le réifier mécaniquement, l'homme se sent maintenant inférieur à la machine intelligente, et chercherait alors activement sa propre réification. Si la réification a d'abord été ressentie comme une honte et un rabaissement, elle est donc devenue un moyen pour contrer ce nouveau degré de la honte, la honte prométhéenne, issue de sa confrontation avec la machine supérieure. C'est le moment où "il approuve sa propre réification et rejette sa non-réification comme un défaut"[13]. C'est exactement la volonté de se faire soi-même machine, de s'aligner, que l'on retrouve au sein du mouvement technologique contemporain.

La question de l'alignement

Si c'est par ce décalage de capacités que survient ce sentiment de honte, l'individu va, pour Anders, tenter de trouver une résolution technologique. Si c'est notre corps qui fait défaut, qui se retrouve limité, borné, alors il est nécessaire de chercher à l'améliorer: "l'homme contemporain cherche à échapper à cette calamité en alignant son corps sur ses instruments grâce au *human engineering*" afin de tendre vers la "parfaite consubstantialité instrumentale"[14]. Aujourd'hui, au sein de ce mouvement technologique, certains individus cherchent à rester dans la course face aux artefacts, notamment en cherchant à se faire soi-même machine

[12] G. Anders, *op. cit*., p. 46.
[13] *Ibid.*
[14] Ivi, p. 53.

et à viser cette consubstantialité. Anders s'inquiétait déjà de ceux qui cherchaient à participer à cette réification, de ceux qui voulaient tenter de s'aligner sur leurs créations. Il voyait dans ce souhait de se réifier, de se faire soi-même machine, le signe ultime de notre soumission au progrès technologique. Aujourd'hui, Musk illustre cette ambition, ce désir de nous aligner avec la machine, à travers Neuralink. Dans une optique de concurrence avec les programmes d'IA, l'entrepreneur cherche à permettre aux individus de s'aligner sur leurs capacités quantitatives de traitement de l'information.

Le mouvement technologique porte aussi l'objectif d'un alignement optimisé mais non invasif de l'individu, via des techniques de manipulation comme le *nudge* directement inspirées des théories du behaviorisme radical. Le behaviorisme radical, s'est notamment développé grâce au psychologue américain B.F Skinner, qui ressentait lui aussi une profonde honte. Honte de nos errements en tant qu'individus, de notre violence, de nos faiblesses. Mais Skinner ne voulait pas recourir à des techniques aussi invasives que Musk pour nous améliorer car il craignait que l'acceptation de ces technologies par la société ne prenne trop de temps. Dans les années 70, il promouvait en revanche la manipulation de nos comportements afin de les optimiser: "En bref, nous devons apporter de vastes changements au comportement humain. Ce dont nous avons besoin, c'est d'une technologie du comportement (*technology of behaviour*)"[15]. Cette manipulation doit s'opérer en observant et en analysant la multitude de processus cognitifs qui gouvernent nos actions. Prendre conscience de ces processus qui sont largement déterminés par des éléments objectifs c'est donc fondamentalement vouloir se positionner par-delà la liberté et la dignité, pour reprendre le titre de son plus grand ouvrage. Liberté et dignité qui lui apparaissent comme des concepts caducs, qui ne servent qu'à cacher notre ignorance des éléments qui nous déterminent réellement: "L'homme autonome ne sert qu'à expliquer les choses que nous ne sommes pas encore en mesure d'expliquer d'une autre manière. Son existence dépend de notre ignorance et il perd naturellement son statut au fur et à mesure que nous en savons plus sur les comportements"[16]. Nous parlons de la liberté d'une action car en réalité nous n'avons pas encore compris tous les phénomènes psychologiques et biochimiques qui ont amené tel individu à la réaliser. Dans l'immense majorité des cas, il deviendrait facile d'orienter les actions des individus en comprenant ces mécanismes et en manipulant les éléments objectifs et extérieurs avec lesquels ils interagissent. C'est tout le souhait de Skinner, de faire du monde

[15] B.F. Skinner, *Beyond freedom and dignity*, coll. "Penguin Books", Repr, Bungay, Suffolk, Clay 1982, p. 10.
[16] Ivi, p. 20.

un grand laboratoire pour créer un environnement performant qui soit à même d'orienter les individus vers les actions qui assurent la pérennité de notre espèce. Ce processus créatif est fondamentalement itératif: chaque modification objective de l'environnement doit être analysée ainsi que la réponse de l'individu à cette modification, qu'elle confirme le résultat attendu ou non. On sera alors en mesure de comprendre quelles sont les actions qui vont renforcer tel ou tel comportement chez un individu. Et on pourra alors développer cette manipulation et l'étendre si elle est efficace et ainsi participer au processus d'évolution de notre espèce. Cette théorie de la manipulation des comportements est à la source des techniques de *nudge*, ces manipulations comportementales indolores qui orientent les individus vers les actions souhaitées par le concepteur du *nudge*. Dans un monde technologique où nous sommes en permanence confrontés à des signaux extérieurs, il est donc possible d'imaginer trouver les *nudges* adéquats pour lutter contre cette honte prométhéenne qui nous habite. Pour nous rendre enfin plus performants, plus efficaces, plus optimaux, mais de façon discrète, sans nous ne contraindre ni nous forcer, via des "mesures non aversives"[17]. La production de données, en ligne mais également dans le monde réel via les objets connectés, rend aujourd'hui possible cet objectif de transformer le monde en un gigantesque laboratoire à ciel ouvert. Car le mouvement technologique actuel et l'ambition qu'il porte d'une captation intégrale du réel et de sa réduction à des données fait que cet idéal n'a jamais été aussi proche. A mesure que nous allons pouvoir collecter et analyser les données il deviendra de plus en plus simple de comprendre les leviers qui orientent les comportements individuels. C'est déjà le cas avec les réseaux sociaux, les applications et autres outils digitaux que nous utilisons au quotidien. Tous sont conçus pour nous garder "accrochés"[18] et ainsi maximiser le temps que nous y passons. Cette maximisation passe déjà par la compréhension des biais cognitifs humains et le recours à des notifications, des couleurs, des sonneries, des rappels... autant d'éléments qui jouent comme des incitations et des renforcements biochimiques pour nous retenir sur ces applications. Skinner évoquait déjà ce rapport technicien aux problèmes sociaux et politiques du monde contemporain: "ce qu'il faut, c'est plus de contrôle, pas moins, et c'est en soi un problème d'ingénierie de première importance"[19]. Il est frappant de voir qu'Anders avait également prévu ce type de velléités de manipulation comportementale afin d'aligner les individus sur la performance des machines: "sans craindre la peine ni les

[17] Ivi, p. 37.
[18] Voir N. Eyal, *Hooked: comment créer un produit ou un service qui ancre des habitudes*, Eyroll, Paris 2018.
[19] B.F. Skinner, *op. cit.*, p. 175.

supplices, il consacre ses efforts et son ingéniosité à effacer de son travail toute trace de spontanéité et d'humanité, puisqu'il espère dépasser les limites fatales de son absence de liberté pour atteindre enfin au souverain bien de l'utilité totale."[20]

Au-delà de la question de la honte et de l'alignement, intéressons-nous maintenant à la seconde thèse d'Anders et à la question de la responsabilité vis à vis de ces machines.

L'impossible responsabilité?

Anders s'appuie sur une anecdote militaire et la démission d'un général américain, désavoué publiquement lorsque sa décision a été remise en question par un "cerveau électrique", pour illustrer sa seconde thèse: "ce que nous produisons excède notre capacité de représentation et notre responsabilité"[21]. Les années 50 voyaient déjà émerger les premières entités capables de collecter, traiter et analyser une quantité surhumaine de données pour produire des prédictions quantitativement meilleures que celles des humains. C'est ainsi que les premières expériences de délégation de la prise de décision ont eu lieu et que certains individus ont "remis le pouvoir de décider à un instrument"[22]. Pour Anders, le levier principal de cette délégation est la méfiance vis-à-vis de l'homme, qui n'a, lui, qu'un cerveau humain. La honte que l'on ressent vis-à-vis de la faillibilité humaine va nous faire progressivement choisir la machine comme entité décisionnaire. Cette question de la délégation est pour Anders problématique car elle se fait pourtant sans que nous soyons en mesure ni de nous représenter ni de rendre compte et d'expliquer cette décision. Et c'est ici qu'Anders propose une analyse visionnaire qui prend tout son sens aujourd'hui, avec les services d'IA. Services dont nous ne sommes, aujourd'hui, pas en mesure de nous représenter le processus de décision les amenant à produire telle recommandation, ni à en rendre compte. Ces deux fonctions, de représentation et de responsabilité, renvoient à deux concepts propres au *machine learning* contemporain[23]: l'interprétabilité et l'explicabilité de la recommandation d'un programme d'IA, en particulier des modèles d'apprentissage non supervisés. D'une part, les concepteurs des programmes ne peuvent pas, aujourd'hui, retracer le parcours d'apprentissage des programmes d'IA ni proposer une inter-

[20] G. Anders, *op. cit*., p. 59.
[21] Ivi, p. 11.
[22] Ivi, p. 79.
[23] G. Leilani, *et al.*, *Explaining Explanations: An Overview of Interpretability of Machine Learning*, in "arXiv":1806.00069, 3 février 2019.

prétation des résultats qu'ils fournissent. D'autre part, et c'est une conséquence directe du manque d'interprétabilité, le programme ne peut pas non plus rendre de compte sur sa recommandation finale. Il ne peut pas l'expliquer de façon claire et en s'appuyant sur une sémantique classique compréhensible par des individus, qu'ils soient des experts ou non. Ce manque d'explicabilité renvoie donc à la question de la responsabilité: car si l'individu délègue la prise de décision à un agent irresponsable, qui ne peut rendre compte de son action et l'expliquer, alors l'enjeu de savoir à qui imputer la responsabilité devient central.

Au-delà de la question de la responsabilité, se pose aussi celle de notre propre capacité à agir. Car si Anders évoquait un monde soumis à une "pathologie collective" qui nous mène au point de "construire un monde au sein duquel nous serons incapables de marcher"[24], il semblerait que notre faculté de penser, d'agir de façon de responsable, soit aussi ici rendue caduque par cette délégation de l'exécution à des services d'IA.

De la honte à la délégation: la place de l'individu dans le mouvement technologique

La honte nous semble apparaître comme le premier temps d'un processus complexe. Cette honte implique la mise en place d'une délégation de la prise de décision à des machines, délégation décrite par Anders et que l'on retrouve avec les services d'IA. Cette délégation de l'homme à la machine est donc la seconde étape de ce processus. La troisième pourrait être une dégradation des compétences des individus, à mesure que la machine prend en charge un nombre croissant d'actions. On voit donc que le constat de la faillibilité de l'individu est à la fois cause et conséquence de cette délégation de la responsabilité: la piètre performance humaine est d'abord cause de la délégation mais est également engendrée par cette délégation qui ne permet pas à l'individu de développer ses compétences. L'individu se retrouve donc dans un cercle vicieux de la délégation qui devient en réalité auto-réalisatrice. C'est le constat que dresse Crawford: "au fur et à mesure que cette boucle progresse, l'hypothèse sous-jacente de l'automatisation de notre incompétence devient progressivement auto-réalisatrice"[25].

Cette inquiétude face à la perte de compétences liée à l'essor des nouvelles technologies est en particulier mis en lumière par Tristan Harris, activiste et ancien employé de Google, à travers le concept de "*human*

[24] G. Anders, *op. cit*., p. 32.
[25] M. Crawford, *Why we drive: on taking back control*, The Bodley Head, Londres 2020, p. 98.

downgrading"[26]: les technologies fondées sur l'économie de l'attention nous enferment dans une spirale qui mine toute forme de réflexion et nous gardent fascinés, loin de toute forme d'émancipation. Harris cherche à promouvoir l'essor d'une technologie alternative qui chercherait à promouvoir d'autres valeurs que la pure performance, performance à laquelle l'individu ne pourra égaler la prouesse de la machine. Harris s'inscrit dans le sillage de la pensée d'Illich, qui promouvait la mise en place d'outils conviviaux, efficaces mais qui ne nient pas l'autonomie et la liberté individuelles: "j'entends par convivialité l'inverse de la productivité. (…). La convivialité est la liberté individuelle réalisée dans la relation de production au sein d'une société dotée d'outils efficaces"[27].

Afin de tendre vers ce développement technologique alternatif, soulignons ce qui fait notre singularité face à ce qu'Ellul nomme le "système expert"[28]. Cela passe en partie par souligner ce que ce système n'est pas en mesure de faire.

Ce qu'il faut tout d'abord relever est que si les processus de prise de décision des services d'IA sont opaques, leur finalité à elle été définie par des humains. Si l'on reprend l'exemple d'Alphago, son action est optimale et efficace, mais en rien autonome. Littéralement, le programme ne se fixe pas son propre *nomos*, sa propre loi qui guide son action. C'est bien les dirigeants de DeepMind qui lui ont assigné cet objectif, l'ont programmé et lui ont fourni puissance de calcul et données pour l'atteindre. Si l'action d'Alphago est bien intentionnelle, *intentio* signifiant tendre vers: et en l'occurence le programme est bien tendu vers sa raison d'être, son objectif de remporter la partie. Ce n'est pas lui qui réalise cette mise en tension, qui fixe cet objectif, ce *telos*.

Au-delà de l'absence de ce *telos*, les programmes d'IA présentent encore différents écueils: ils nécessitent une immense quantité de données pour atteindre le niveau d'un humain (i); ils ne sont pas robustes face à des attaques basiques qu'un humain déjouerait facilement (*adversarial attacks*) (ii); ils ne sont pas en mesure de transférer leur apprentissage dans un contexte différent (iii); ils ne font pas preuve de bon sens (iv)[29].

En considérant ces difficultés techniques non encore résolus ainsi que

[26] T. Harris, *Opinion | Our Brains Are No Match for Our Technology*, in "The New York Times", rubrique "Opinion", 5 décembre 2019 (en ligne: https://www.nytimes.com/2019/12/05/opinion/digital-technology-brain.html).
[27] I. Illich, *La convivialité*, Éd. Points, Paris 2014, p. 28.
[28] Voir le § "La rationalité" dans J. Ellul, *Le Bluff technologique*, Pluriel, Paris 2014, pp. 301-321.
[29] University College London et DeepMind, Deep Learning Lectures | 10/12 | Unsupervised Representation Learning, 22 juin 2020 (en ligne: https://www.youtube.com/watch?v=f0s-uvvXvWg).

ce que nous avons eu jusqu'ici, nous pouvons donc synthétiser trois éléments essentiels à propos de ces services:

– ils ne sont pas autonomes dans la définition de leur *telos;*

– ils possèdent de nombreuses faiblesses comparées à l'intelligence humaine;

– ils sont irresponsables et ne peuvent rendre compte de leurs actions.

Le paradoxe fondamental de ces limitations est qu'elles s'enracinent dans la condition technologique propre aux services d'IA et en particulier le fait que ces programmes ne sont pas dotés de corps biologique.

Le corps est en effet le siège du moteur essentiel de toute motivation humaine, de toute quête, de toute recherche, de toute poursuite d'objectifs: ce moteur, c'est l'homéostasie. Concept physiologique mis notamment en avant par le neuroscientifique Antonio Damasio[30]. L'homéostasie, c'est le rééquilibrage dynamique permanent de nos affects et de nos sentiments, qu'ils soient positifs ou négatifs, épanouissement ou souffrance, et qui nous poussent à agir. Créer une dichotomie stricte entre des processus intellectuels et des processus corporels, qui seraient séparés, est donc illusoire. Cette cohabitation se fait notamment par l'intermédiaire des sentiments qui sont, "à tout point de vue – de manière simultanée et interactive – des phénomènes liés à la fois au corps et aux systèmes nerveux"[31]. Et ces sentiments font varier l'équilibre interne de l'organisme, ce qui nous pousse à l'action, à la résolution de ce déséquilibre. Sans les sentiments, il serait pour Damasio "impossible d'évoquer la pensée, l'intelligence et la créativité"[32]. Ils influencent nos décisions, imprègnent tous les aspects de notre existence "et sont à la source de l'élaboration des pratiques et des instruments culturels humains"[33]. La construction de notre esprit est le fruit de cette coopération entre le cerveau et l'organisme, via les sentiments. Et c'est cette coopération qui nous pousse à agir, à régler nos déséquilibres, jusqu'à la fondation de nos communautés et nos sociétés. Damasio insiste sur le fait que l'IA pourra très bien exceller dans les domaines analytiques, il cite d'ailleurs précisément le jeu de go, sans pouvoir jamais être motivé et posséder son principe de mise en mouvement propre. Il serait donc toujours l'instrument dans les mains d'un créateur dont dépendrait sa raison d'être. Dépourvue de sentiments, l'IA ne pourrait pas éprouver ce déséquilibre dynamique. Elle ne pourrait pas être amenée à ressentir ces variations et être poussée à l'action, à percevoir des émotions, à se sentir tantôt apaisée, tantôt vulnérable, tantôt joyeuse ou tantôt désespérée.

[30] A. Damasio, *L'ordre étrange des choses*, Odile Jacob, Paris 2017.
[31] Ivi, p. 180.
[32] Ivi, p. 200.
[33] *Ibid.*

Si tout cela passe par le corps, le réductionnisme du mouvement technologique est un matérialisme fonctionnaliste qui pense pouvoir copier le corps, l'assimilant à un système mécanique doté de fonctions particulières qu'il serait possible d'isoler et de répliquer (en particulier au sein du cerveau). Cette volonté de réduire le corps à un système mécanique, d'assimiler l'homme à une machine, est symptomatique de notre époque cybernétique, et pour reprendre les mots de Castoriadis, nous sommes parvenus à "un degré d'enfoncement dans l'imaginaire (et) aucune société primitive n'a jamais appliqué aussi radicalement les conséquences de ses assimilations des hommes à autre chose, que ne l'a fait l'industrie moderne de sa métaphore de l'homme automate"[34]. Cette volonté de réduction est aujourd'hui portée par le mouvement technologique et renforcée par la sophistication des instruments, notamment en imagerie cérébrale, pour mettre en lumière le fonctionnement du cerveau. Ainsi, il est par exemple maintenant objectivement établi la différence entre la perception et la conscience d'un objet, car nous percevons davantage de phénomènes que nous en avons réellement conscience[35]. Néanmoins, au sujet de la conscience c'est un tort de réduire la conscience à cette seule expérience objective "d'avoir conscience de", comme peut le faire Musk. Il existe en effet une nuance fondamentale entre l'expérience objective de la conscience et l'acte réflexif et subjectif de prise de conscience. Michel Bitbol, philosophe et directeur de recherche émérite aux Archives Husserl explique qu'au mieux, la technologie permettra "d'observer les corrélats neuronaux d'activités mentales"[36] mais jamais il ne pourra être possible de rendre quantitativement compte de "la composante subjective de la conscience, de la qualité vécue de l'expérience, du ressenti lui-même". Cette expérience phénoménale subjective semble donc être bien l'un des éléments qui nous permettent de définir l'irréductibilité de notre nature humaine, d'une composante qui ne peut être réduite à une donnée quantifiable. Pourtant, le mouvement technologique cherche à percer cela, comme le prédisait Skinner pour la pensée: "le dernier bastion (*stronghold*) de l'homme autonome est peut-être cette activité cognitive complexe qu'est la pensée"[37].

Le philosophe britannique Pierre Hacker reprend cette thématique de la singularité de l'espèce humaine dans l'un des discours qui constituent son dernier ouvrage: la cause formelle aristotélicienne de l'homme est sa *psyche* et on ne peut pas chercher à la réduire à une substance:

[34] C. Castoriadis, *L' institution imaginaire de la société*, Éd. du Seuil, coll. "Points Essais", nº 383, Paris 2006, p. 238.
[35] Dossier *Révéler la conscience,* in "La Recherche", 565, avril 2021, pp. 18-59.
[36] Ivi, pp. 20-23.
[37] B.F. Skinner, *Beyond freedom and dignity*, cit., p. 188.

> Nous ne devrions pas penser la *psyche* comme à un être mais plutôt comme la forme des êtres vivants. La *psyche* n'est pas incarnée, c'est plutôt l'organisme vivant qui est *empsuchos*, animé. La *psyche* est constituée par les différents pouvoirs qui informent les êtres vivants et en vertu desquels ils sont le genre d'êtres qu'ils sont[38].

L'ambition réductionniste de voir "l'esprit comme une boite noire recevant des inputs de nos organes sensoriels et produisant des outputs comportementaux"[39] avec "l'esprit (qui) joue le rôle d'intermédiaire causal entre ces entrées et sorties au mode, d'états mentaux conçus comme des états fonctionnels d'un organisme entretenant des relations causales" est donc illusoire. On ne peut pas réduire les individus à des fonctions que l'on pourrait imiter à travers des machines car "cette conception de l'esprit et du mental ne fait manifestement aucune place pour la dimension qualitative des expériences qu'éprouvent les créatures dotées de sensibilité". Les individus ne sont donc pas les simples possesseurs de fonctions comme "penser" ou "ressentir" et nous sommes uniquement dotés d'aptitudes qui sont irrémédiablement liées à cette *psyche* qui est à la fois nutritive, sensitive et rationnelle. La volonté de réduire ces aptitudes à une fonction précise d'une partie du corps représente un "paralogisme méréologique"[40]: "lorsqu'on attribue à une partie d'une chose des propriétés qui ne peuvent être attribuées qu'à cette chose prise en son ensemble". Il est donc illusoire de croire qu'il pourrait être possible de réaliser une machine qui penserait comme un cerveau car ce n'est pas le cerveau qui pense, ni d'ailleurs est en possession de nos pensées. Penser est une aptitude propre de l'humain doté de sa *psyche*. Les machines ne sont donc "ni conscientes ni inconscientes. Elles n'ont pas de plaisir à ce qu'elles font, pas plus qu'elles ne souffrent. Elles n'éprouvent ni amour ni haine. Elles ne délibèrent pas sur la ligne de conduite à adopter (…) elles ne connaissent pas la différence entre le bien et le mal"[41].

S'éprouver à notre condition corporelle limitée serait donc à la fois cause de cette honte mais proposerait une piste pour en sortir, alors que viser l'efficace et l'optimal, le rationnel et l'immédiat, se trouverait à l'opposé de cette pratique. Anders avait déjà saisi les difficultés à admettre cette nécessité de mener un combat propre à notre singularité d'individus dotés d'aptitudes spirituelles: "rien ne nous caractérise davantage, nous, les hommes d'aujourd'hui, que notre incapacité à rester spirituellement

[38] P. Hacker, *Dialogues sur la pensée, l'esprit, le corps et la conscience*, Le Kremlin-Bicêtre, Marseille 2021, p. 36.
[39] Ivi, pp. 20-21.
[40] Ivi, p. 31.
[41] Ivi, p. 22.

up to date par rapport au progrès de notre production"[42]. Ellul estimait également que dans ce mouvement qui s'accélère "la tragédie intellectuelle et culturelle (...), c'est que nous sommes dans un milieu technicien qui ne permet plus la réflexion"[43].

Bibliographie

Anders G., *L'obsolescence de l'homme: sur l'âme à l'époque de la deuxième révolution industrielle (1956)*, Éd. de l'Encyclopédie des nuisances, Paris 2002.

Castoriadis C., *L' institution imaginaire de la société*, Éd. du Seuil, coll. "Points Essais", 383, Paris 2006.

Crawford M., *Why we drive: on taking back control.*, The Bodley Head, Londres 2020.

Illich I., *La convivialité*, Éd. Points, Paris 2014.

Skinner B. F., *Beyond freedom and dignity*, Repr, Bungay, Suffolk, coll. "Penguin Books", Clay 1982.

Gilpin L. H., et al., *Explaining Explanations: An Overview of Interpretability of Machine Learning*, in "arXiv:1806.00069" [cs, stat], 3 février 2019 (en ligne : http://arxiv.org/abs/1806.00069). ArXiv: 1806.00069.

Harris T., "Opinion | Our Brains Are No Match for Our Technology", *The New York Times*, rubrique "Opinion", 5 décembre 2019 (en ligne: https://www.nytimes.com/2019/12/05/opinion/digital-technology-brain.html).

Silver D., et al., *Mastering the game of Go without human knowledge*, in "*Nature*", 550, 7676, octobre 2017, pp. 354-359.

Swisher K., *Opinion | Elon Musk: A.I. Doesn't Need to Hate Us to Destroy Us*, in "The New York Times", rubrique "Opinion", 28 septembre 2020 (en ligne: https://www.nytimes.com/2020/09/28/opinion/sway-kara-swisher-elon-musk.html).

Thompson N., *Tristan Harris: Tech Is 'Downgrading Humans. It's Time to Fight Back*, in "Wired", sans date (en ligne: https://www.wired.com/story/tristan-harris-tech-is-downgrading-humans-time-to-fight-back/ ; consulté le 30 septembre 2021).

University College London et DeepMind, *Deep Learning Lectures | 10/12 | Unsupervised Representation Learning*, 22 juin 2020 (en ligne: https://www.youtube.com/watch?v=f0s-uvvXvWg).

[42] G. Anders, *op. cit*., p. 30.
[43] E. Jacques, *Le Bluff technologique*, Pluriel, Paris 2014, p. 277.