PHARES

Vol. XIX nº 2 Automne 2019

Revue philosophique étudiante

Nous remercions nos partenaires :

- La Faculté de philosophie de l'Université Laval
- L'Association des étudiantes et des étudiants de Laval inscrits aux études supérieures (AELIÉS)
- L'Association des chercheur(e)s étudiant(e)s en philosophie de l'Université Laval (ACEP)
- La Confédération des associations d'étudiantes et d'étudiants de l'Université Laval (CADEUL)
- La Société de philosophie du Québec (SPQ)
- La Fondation de l'Université Laval
- L'Association générale des étudiantes et étudiants prégraduéEs en philosophie de l'Université Laval (AGEEPP)
- La Coop Zone

Revue *Phares*
Bureau 514
Pavillon Félix-Antoine-Savard
Université Laval, Québec
G1K 7P4

revue.phares@fp.ulaval.ca
www.revuephares.com

ISSN 1496-8533

PHARES

Comme son nom l'indique, la revue *Phares* essaie de porter quelques lumières sur l'obscur et redoutable océan philosophique. Sans prétendre offrir des réponses aptes à guider ou à éclaircir la navigation en philosophie, cette revue vise, en soulevant des questions et des problèmes, à signaler certaines voies fécondes à l'exploration et à mettre en garde contre les récifs susceptibles de conduire à un naufrage. En outre, le pluriel de *Phares* montre que cette revue entend évoluer dans un cadre aussi varié et contrasté que possible. D'une part, le contenu de la revue est formé d'approches et d'éclaircissements multiples : chaque numéro comporte d'abord un ou plusieurs Dossiers, dans lesquels une question philosophique est abordée sous différents angles. Le dossier publié à la session d'hiver est accompagné d'une section Varia, qui regroupe des textes d'analyse, des comptes rendus, des essais, etc. Finalement, nos numéros comportent parfois une section Répliques dans laquelle il est possible de répondre à un texte précédemment publié ou d'en approfondir la problématique. D'autre part, la revue *Phares* se veut un espace d'échanges ouverts à tous les étudiants et étudiantes intéressé.e.s par la philosophie. Pour participer aux prochains numéros, voir la politique éditoriale publiée à la fin du présent numéro.

Nous vous invitons à consulter notre site Internet (www.revuephares.com), où vous aurez accès à tous les articles parus dans *Phares*.

Table des matières

Dossier :

« Philosophie et éthique
de l'intelligence artificielle »

Introduction – Les réflexions philosophiques étudiantes sur l'intelligence artificielle

Keven Bisson, *Codirecteur de la revue Phares, Université Laval et* Romane Marcotte, *Codirectrice de la revue Phares, Université Laval*
Avec les commentaires de Jocelyn Maclure*, professeur titulaire à la Faculté de philosophie de l'Université Laval*

Au cours de la dernière décennie, les développements fulgurants de l'intelligence artificielle (IA) ont pris une place centrale dans le monde intellectuel. Le Québec est d'ailleurs un acteur privilégié de ces développements et des réflexions qui y sont liées : le Québécois Yoshua Bengio a gagné en 2018 l'un des plus importants prix dans le domaine pour son travail sur les réseaux neuronaux profonds (le prix Turing), des entreprises comme Google et Facebook viennent s'installer à Montréal, l'Observatoire international sur les impacts sociétaux de l'IA et du numérique s'est installé à Québec, etc. La Faculté de philosophie de l'Université Laval ne fait pas exception, et a augmenté considérablement son offre de cours touchant le sujet de l'intelligence artificielle. C'est dans ce contexte que s'inscrit ce numéro de la revue Phares, qui présente un aperçu du travail des étudiantes et étudiants en philosophie sur ce sujet, afin de stimuler les réflexions présentes et à venir sur ce thème. Parmi les multiples inquiétudes que soulève le développement effréné des algorithmes de l'IA, deux grandes problématiques intéressent les recherches philosophiques présentées dans ce dossier : la possibilité d'une conscience artificielle et les conséquences éthiques de l'utilisation et de la constitution de machines intelligentes.

D'abord, concernant la première problématique, plusieurs chercheur.es croient que la création d'une intelligence artificielle consciente est possible. Toutefois, le fait que l'être humain ait

des émotions et un corps, qui sont partie intégrante de son expérience consciente, semble être un obstacle important à l'atteinte de cet objectif. Est-il possible de créer une IA qui dépasserait cette difficulté, c'est-à-dire qui serait dotée d'un point de vue subjectif sur le monde qui soit comparable au nôtre ? C'est à cette question que s'attardera le premier texte de notre dossier. Cet article est avant tout un refus de réduire la conscience humaine à une somme de processus neuronaux, de penser notre esprit comme un simple logiciel implanté dans la matière du cerveau - comme le font certaines théories cognitivistes sur lesquelles s'appuient les tenants de l'IA forte. Lorsqu'on aborde la conscience sous un autre angle, c'est-à-dire en la considérant comme le lieu de nos vécus des phénomènes du monde, la possibilité qu'une machine puissent un jour posséder une conscience comparable à la nôtre se voit considérablement réduite. Pour sa démonstration, Geneviève Fréchette détaille d'abord les caractéristiques qui font de notre conscience une conscience phénoménale en s'appuyant sur la phénoménologie de Husserl. Les principales caractéristiques retenues seront entre autres : l'ancrage dans un corps matériel, la possession d'une volonté, la possession d'un point de vue subjectif, la capacité d'expérimenter des vécus intentionnels et celle d'opérer une synthèse des vécus. L'auteure en vient finalement à déterminer l'impossibilité pour une IA forte de posséder une conscience qui puisse être dite phénoménale, puisqu'elle serait incapable d'appréhender une expérience depuis un point de vue proprement subjectif, relevant du domaine du sens.

Si on considère cependant possible la création d'une IA consciente, un autre problème se pose : jusqu'à maintenant, l'être humain légitimait son statut moral particulier par la possession d'une conscience. Comment alors considérer moralement une entité artificielle qui possèderait une intelligence égale - voire supérieure - et une conscience semblable à celle de l'être humain ? Cette question est particulièrement importante puisqu'on envisage la conception de robots qui entreraient en relation avec des êtres humains, notamment pour accomplir à leur égard diverses tâches de soin. Dans cette dernière optique, on songe notamment à créer des sexbots, c'est-à-dire des robots conçus pour avoir des relations

intimes et sexuelles avec les êtres humains. La sexualité constitue déjà un aspect de la vie humaine plein de nuances et de paradoxes. La possibilité de la conception de tels robots ainsi que les implications éthiques de leur mise en marché nécessitent donc une réflexion approfondie, ce à quoi s'attèle le second texte de notre dossier. Dans cet article, Samuel Nepton confronte directement l'affirmation de Joanna Bryson selon laquelle les robots devraient être considérés comme des esclaves et traités comme de simples objets. Selon lui, que l'on songe à la construction de sexbot « faibles » (des robots qui ne feraient qu'imiter des sentiments), ou de sexbots « forts » (des robots véritablement sensibles et conscients d'eux-mêmes), ces machines constitueraient un véritable angle mort aux thèses de Bryson. Cet angle mort se décline en deux catégories. D'un côté, la réduction des sexbots au statut d'objet risquerait d'aggraver la situation de vulnérabilité des humains avec lesquels ils doivent entretenir une relation de soin. De l'autre côté, dans l'optique où ces robots seraient conçus comme des êtres sensibles, ils seraient susceptibles de souffrir et donc auraient besoin de protection.

L'auteur du texte nous fera d'abord remarquer l'importance que ces machines ne soient pas de simples jouets sophistiqués de masturbation, puisque leur conception naît d'un besoin que certaines personnes ont de nouer de véritables relations. C'est donc au niveau relationnel qu'il faudra juger le statut de ces robots - une approche possible grâce au travail de Mark Coeckelberg. Une telle approche permet de dégager trois manières dont la réduction du robot au stade d'objet rendrait les « propriétaires » des sexbots, et les sexbots eux-mêmes, vulnérables. D'abord, certains humains qui seraient incapables de trouver l'amour, souvent des personnes marginalisées, seraient encore plus isolés socialement du fait qu'on les considèrerait en relation avec de simples « objets ». Dans un tout autre ordre d'idée, certains sexbots personnalisés pourraient permettre à des êtres humains au comportement sexuel déviant (on nomme par exemple le cas de la pédophilie) d'exacerber leurs tendances pathologiques. Finalement, dans le cas où l'on concevrait des sexbots forts, ceux-ci seraient capables d'expérimenter des sentiments comme l'amour et

la tendresse, mais seraient aussi sujet à la souffrance. Pour contrer ces problèmes, l'auteur propose l'adoption d'un cadre légal minimal.

L'utilisation actuelle de l'IA concerne également des aspects plus près de notre réalité, comme l'augmentation de l'efficacité de nos institutions par le traitement de données massives orientant la prise de décision. L'IA traite en effet déjà des demandes de libérations conditionnelles, ou encore de l'octroi de prêts dans certaines banques. L'IA présente deux avantages par rapport à l'être humain pour accomplir ce type de tâches. D'un côté, elle est plus « intelligente » que l'être humain, puisqu'elle peut considérer plus de facteurs complexes en même temps pour prendre une décision. D'un autre côté, le caractère « artificiel » de cette intelligence s'oppose au caractère « naturel » de l'intelligence humaine, teintée d'émotions, d'un vécu particulier et de préjugés. Cependant, contrairement à ce qu'on pouvait espérer, la courte expérience que nous avons de cette utilisation de l'IA montre que les algorithmes sont également empreints de discriminations. Dans le troisième texte du dossier, Sandrine Charbonneau s'attarde à cette discrimination algorithmique. En plus de nous montrer par plusieurs exemples l'ampleur des conséquences que ces algorithmes biaisés peuvent avoir sur la vie des individus, l'auteure explore les diverses causes qui pourraient être à leur origine, notamment le manque de diversité et de sensibilisation chez les programmeurs et programmeuses. À ce problème s'ajoute celui de l'opacité des algorithmes de deep learning, dont le fonctionnement est considéré comme un secret d'affaires. En effet, les recherches qui souhaiteraient repérer les articulations problématiques de ces algorithmes sont difficiles à effectuer en raison de cette impossibilité à accéder aux algorithmes. En réponse à ces problématiques, l'article propose plusieurs pistes d'encadrement d'utilisation des algorithmes de deep learning, notamment la mise en place d'une procédure rigoureuse d'audit afin de tester ces programmes avant de les lancer sur le marché. Il souligne aussi l'importance de sensibiliser les programmeurs et programmeuses responsables des collectes de données à la source du fonctionnement des algorithmes, afin qu'ils et elles soient conscient.es de l'impact de leurs biais dans leur travail. Finalement,

l'article nous met en garde contre une confiance aveugle dans ces programmes d'intelligence artificielle, et insiste pour que la décision finale reste prise par un humain.

Les algorithmes ne se restreignent pas seulement à l'accomplissement de tâches que nous faisions, ils peuvent en accomplir de nouvelles que nous n'aurions jamais pensé possibles si elles devaient être faites par des êtres humains. En effet, les algorithmes permettent de sélectionner le contenu présenté sur différentes plateformes publiques afin de le rendre plus personnalisé pour son utilisateur ou son utilisatrice. Or, l'organisation du contenu de ces plateformes très fréquentées (pensons aux réseaux sociaux), n'est pas sans influencer notre manière de voir le monde, et notamment nos opinions politiques. Le dernier texte du dossier explore l'impact de cette influence dans le contexte de la polarisation des opinions politiques aux États-Unis. En effet, Éric Gagnon remarque que le fossé idéologique entre partisans démocrates et républicains n'a cessé de se creuser au cours des dernières années, et constate que la plateforme Facebook a été un des milieux à l'origine de cette division. Afin de développer le lien entre les algorithmes de ce réseau social et la progressive radicalisation des opinions politiques américaines, l'article explore le fonctionnement de la raison humaine. En suivant les thèses d'Hugo Mercier et Dan Sperber, Éric Gagnon se trouve en mesure de déterminer que les environnements les plus propices à l'exercice de la raison ainsi qu'au déroulement d'une délibération saine sont avant tout des lieux dits « hétérogènes ». Ces milieux se caractérisent avant tout par une pluralité d'opinions, qui ne permettent pas à l'individu de se conforter dans ses propres croyances, et qui le poussent à formuler des arguments rigoureux. Dans un contexte contraire, l'homogénéité d'un milieu rend la raison paresseuse, et, en raison du biais du parti pris, moins apte à produire des arguments solides. Sachant cela, l'article fait remarquer que les algorithmes utilisés sur les plateformes des réseaux sociaux sont précisément conçus pour nous exposer à du contenu politique adapté à nos préférences. Ces plateformes, telles qu'elles sont utilisées jusqu'à maintenant, ne constituent donc pas des lieux de délibération, mais de radicalisation, et leur fonctionnement

nuit au développement critique des citoyens. On souligne d'ailleurs dans le texte les liens statistiques entre l'utilisation de Facebook de certains élus et la radicalisation de leurs partisans. À partir de cette conclusion, le texte propose plusieurs solutions d'encadrement de l'utilisation des algorithmes sur les réseaux sociaux. Il prône notamment la mise en place de mesures gouvernementales qui exigeraient de ces programmes qu'ils nous présentent un contenu politique aussi diversifié que rigoureux. En attendant de telles réformes, l'auteur de l'article invite son lectorat à diversifier par lui-même le flux d'informations auquel il est confronté au quotidien.

Une conscience phénoménale est-elle possible chez des agents artificiels ?

Geneviève Fréchette, *Université Laval*

RÉSUMÉ : Plusieurs chercheurs dans le domaine de l'intelligence artificielle veulent montrer qu'il est possible de concevoir des machines capables de reproduire le comportement humain. Certains d'entre eux vont encore plus loin et postulent que ces agents artificiels pourraient être munis d'une conscience comme la nôtre - c'est-à-dire d'une conscience phénoménale - et éprouver des sensations, avoir des expériences perceptives, ressentir des émotions, etc. Toutefois, cette thèse est-elle vraisemblable ? Dans cet article, j'analyserai la possibilité d'une conscience phénoménale chez les agents artificiels. Je mettrai en lumière ce qui me paraît être des caractéristiques essentielles d'une telle conscience, pour examiner ensuite s'il est envisageable que celles-ci se retrouvent chez les agents artificiels. Je répondrai finalement que le projet de l'intelligence artificielle forte ne me semble pas réalisable, puisqu'à première vue, un robot ne peut remplir les critères nécessaires d'une conscience phénoménale, qui exige une vie subjective que les agents artificiels ne peuvent posséder.

Introduction

Face aux progrès des sciences cognitives, plusieurs questions d'envergure émergent, notamment en ce qui a trait à l'intelligence artificielle et, plus précisément, à l'intelligence artificielle forte. Selon les défenseurs de l'intelligence artificielle forte, les processus mentaux doivent être compris comme des processus computationnels. Ainsi, il serait possible de concevoir une machine réellement consciente qui serait dotée d'une vie mentale telle que la nôtre. Il va sans dire que si l'idée d'un agent artificiel conscient était réalisable, cela bouleverserait considérablement le paradigme en place. La validation d'une telle hypothèse aurait, entre autres, pour conséquence d'amener la nécessité de repenser la sphère de la politique, de l'éthique, du

droit, de l'éducation, etc. Toutefois, plutôt que de m'écarter vers des conclusions hâtives, je désire porter un regard en amont, en posant une question qui vise l'enjeu même du problème : une conscience phénoménale est-elle possible chez des agents artificiels ? Dans cet article, je tenterai de montrer en quoi cette idée me paraît inadmissible. Pour ce faire, je mettrai en lumière ce qui semble être les critères d'une conscience phénoménale en m'inspirant principalement de la phénoménologie de Husserl. Par la suite, je me pencherai sur les prétentions de l'intelligence artificielle forte et tenterai de montrer que celles-ci sont difficiles à soutenir en raison des caractéristiques propres aux agents artificiels. En effet, ceux-ci ne semblent pas satisfaire aux conditions de possibilité d'une conscience *phénoménale*[1], ce qui me pousse à douter de la validité des thèses de l'intelligence artificielle forte.

1. Qu'est-ce qu'une conscience phénoménale ?

Tout d'abord, il est nécessaire de bien comprendre les termes concernés par l'argumentation qui suivra, en commençant par le concept de conscience phénoménale. Étant donné la nature mystérieuse de la conscience, un éventail de significations lui est attribué, mais je me contenterai de mettre en lumière deux niveaux de définition, soit la conscience d'un point de vue biologique et la conscience d'un point de vue phénoménal. D'un côté, la conscience biologique réfère davantage aux processus mentaux comme la perception, la mémoire, l'imagination, etc. On peut également l'appeler cognition, comprise comme l'ensemble des processus de traitement de l'information. D'un autre côté, la conscience phénoménale - nommée aussi *expérience consciente* - se qualifie plutôt comme l'ensemble des vécus d'un sujet, son expérience subjective, ses qualia[2]. Autrement dit, quelle que soit l'expérience que je fais, la conscience phénoménale est *l'effet* que cela me fait de faire une telle expérience.

2. Les conditions de possibilité de la conscience phénoménale

Afin de déterminer si une conscience phénoménale est possible chez les agents artificiels, il importe d'identifier ce qui paraît être

les conditions de possibilité d'une telle conscience. Dans la mesure où le discours traitant de la conscience se veut être cohérent, quels seraient ses caractéristiques essentielles et ses critères d'effectivité ? Tout d'abord, il semble que toute conscience phénoménale doit avoir un *corps matériel* comme support, en plus d'être dotée d'une libre *volonté*. Ensuite, pour être phénoménale, la conscience doit avoir le caractère de *subjectivité* et d'*intentionnalité*. L'expérience consciente doit également pouvoir faire une *synthèse de ses vécus* et avoir le caractère d'*auto-affection*, c'est-à-dire qu'elle a comme caractéristique d'être à la fois sujet et objet. Enfin, il semble qu'une conscience phénoménale ne peut être pensée qu'avec une *temporalité* et une *spatialité* - au sens où elle est toujours un « ici » et un « maintenant ». Ces caractéristiques seront développées davantage dans les paragraphes qui suivent.

2.1. Le corps matériel

Une conscience phénoménale n'est pas, autant que l'on puisse en juger, imaginable sans un corps. Les deux forment un tout dont le corps est compris comme l'extension de la conscience. Si je suis en mesure de dire que l'expérience que je fais du monde qui m'entoure est toujours *mienne*, c'est grâce à *mon* corps. Je suis capable de toucher mon corps en tant que substrat matériel en même temps que mon corps *se sent* touché ; « d'une part, il est chose physique, *matière*, il a son extension dans laquelle entrent ses propriétés réales[3], la coloration, le lisse, le dur, la chaleur et toutes les autres propriétés matérielles du même genre ; d'autre part, je trouve en lui et je *ressens* "sur" lui et "en" lui : la chaleur du dos de la main, le froid aux pieds[4] ». Le corps est le support de ma conscience du fait qu'il a les organes sensoriels qui me permettent d'être en relation avec le monde, mais aussi parce qu'il est le lieu d'apparition des phénomènes. En ce sens, je suis également convaincue que la conscience phénoménale doit avoir comme support un corps *biologique*[5]. John Searle dira que « la conscience est causée par des processus neuronaux de niveau inférieur dans le cerveau, et est elle-même une caractéristique du cerveau. […] Nous pouvons penser qu'il s'agit d'une "propriété émergente" du cerveau[6] ». Ainsi, la conscience

pourrait être explicable causalement par les propriétés neuronales, mais elle n'est pas pour autant réductible à la somme de celles-ci. Même si nous pouvons penser qu'il y a une relation causale entre les processus neuronaux et l'apparition de la conscience, prétendre que la conscience phénoménale puisse être comprise comme la somme des processus neuronaux est une conclusion illégitime. Autrement dit, le fait que la conscience émergerait du cerveau ne justifie pas à elle seule une thèse, comme celle du computationnalisme fort, qui réduirait la conscience phénoménale à un simple système de traitement de l'information. Or, nous pouvons très bien imaginer que l'expérience consciente est rendue possible grâce au cerveau sans pour autant prétendre que la conscience se résume aux processus neuronaux.

2.2. La subjectivité

L'un des traits les plus importants de la conscience est son caractère subjectif. En effet, il semble qu'une part - voire une très grande partie - de notre expérience échappe à l'objectivité. Par exemple, mon ami reçoit une bonne nouvelle qui le met dans un état de joie intense. Même s'il m'explique ce qui le rend joyeux et qu'il me décrit le plus précisément possible ses émotions, il demeure, pour moi, impossible de saisir l'essence de son vécu. L'expérience vécue est la *sienne*. Or, je peux avoir de l'empathie, imaginer comment il se sent, mais ce sera encore et toujours incomplet. De plus, même si je parvenais à comprendre parfaitement ce qu'il vit, je n'aurai pas accès à l'effet que cela *lui* fait, mais plutôt à ce que cela *me* fait, en tant que *mon* propre vécu. En somme, l'expérience de la conscience phénoménale a comme trait d'essence d'être subjective, elle est toujours pour une conscience unique, de telle sorte que nos expériences sont vécues selon notre point de vue individuel. Nagel dira qu'« il est difficile de comprendre ce que pourrait signifier le caractère *objectif* d'une expérience, indépendamment du point de vue particulier à partir duquel son sujet l'appréhende. Après tout, que resterait-il de l'effet que cela fait d'être une chauve-souris si l'on ôtait le point de vue de la chauve-souris[7] ? » Ainsi, l'entreprise scientifique qui veut expliquer le fonctionnement du cerveau par les

voies de la physique ne parviendrait pas à expliquer la réalité d'un autre ordre, soit celle de la phénoménalité.

2.2.1. Les états intentionnels

De plus, il importe de mentionner que l'expérience consciente n'a de sens qu'au sein d'une compréhension de la conscience comme essentiellement *intentionnelle*. Ainsi, les choses apparaissent à une conscience à travers un vécu et celui-ci est intentionnel dans la mesure où il est toujours un vécu *de quelque chose*. Nous n'aurions accès à rien de ce qui nous entoure si les objets ne nous apparaissaient pas en tant que phénomènes. Or, mon accès à l'arbre devant moi n'est possible que dans la mesure où ma conscience le vise (j'ai conscience *de* l'arbre) et qu'il se phénoménalise à moi (l'arbre apparaît à ma conscience et j'ai un vécu de l'arbre). Je ne parviens jamais véritablement à la chose dans sa réalité empirique, elle n'apparaît à moi qu'en tant qu'objet de signification. Par le moyen du sens, je perçois l'arbre en tant qu'arbre et celui-ci apparaît en tant que *mon* phénomène[8].

2.2.2. La synthèse des vécus de conscience

Cependant, le sens intégral de mon objet de perception ne m'est jamais donné du simple fait que je l'ai visé. Par exemple, quand je regarde l'arbre, je ne peux pas percevoir en même temps toutes ses faces (de côté, de l'intérieur, du dessus, etc.). Ainsi, la conscience doit procéder à une perpétuelle synthèse en vue de constituer l'unité de sens de l'objet et elle opère constamment une liaison entre les vécus de conscience en vue de parvenir à une unité[9]. Par exemple, prenons une personne qui n'a jamais vu ou entendu parler de la neige. Lorsqu'elle en fera l'expérience, elle procèdera à une synthèse de ses vécus, ce qui lui permettra de l'identifier comme froide, blanche, etc. Certes, sa synthèse du vécu de la neige n'aura pas la même complétude que la mienne - habitant au Québec depuis 28 ans - mais, dans les deux cas, la liaison de nos propres vécus antérieurs aura pour effet d'affecter notre expérience phénoménale de la neige.

2.3. *L'auto-affection*

L'accès à une conscience phénoménale requiert un autre critère, celui de l'*auto-affection*. Avoir une conscience phénoménale nécessite d'être en rapport avec les choses du monde. Or, ce n'est pas quelque chose dont nous doutons. Par exemple, nous savons que la neige est froide puisque nous en avons fait l'expérience et si nous pouvons faire une telle expérience, c'est grâce à notre caractère d'auto-affection. Autrement dit, je suis toujours sujet et objet de ma propre expérience. Je ne peux pas me considérer simplement comme un corps ou comme n'importe quel objet biologique. En vérité, l'expérience vécue de mon propre corps révèle, de surcroît, mon expérience à la première personne, moi, en tant que sujet. Ainsi, en parvenant à l'unité de mon propre corps et de son expérience subjective, je confirme ma participation au monde. Le monde est alors toujours *mon* monde ; « le monde environnant n'est pas monde "en soi", mais monde "pour moi", c'est-à-dire justement monde environnant de *son propre* sujet égologique, monde dont le sujet fait l'expérience […] qu'il pose au sein de ses vécus intentionnels avec la teneur de sens qui y est chaque fois impliquée[10] ». On peut dire que la conscience phénoménale affecte le monde en tant qu'elle vit au sein de *ses* objectivations ; les objets du monde sont ses objets. Mais elle est également toujours affectée et transformée par eux dans la mesure où ses comportements sont déterminés par le monde[11].

2.4. *La spatialité et la temporalité*

Cette capacité à avoir une vision unitaire de soi en tant qu'objet et sujet me permet, en outre, de me situer dans le monde en tant que « point zéro ». Les choses du monde entretiennent un rapport à ma conscience en tant qu'elle est le lieu d'apparition des phénomènes. Je ne pourrai jamais m'éloigner de moi, je suis le « ici et maintenant ». En cela, je peux dire que les choses m'apparaissent dans un temps et un espace qui me sont propres, qui sont toujours relatifs à moi. De plus, ma temporalité se dégage du fait que mon expérience consciente est déterminée par mes vécus antérieurs. En reprenant l'exemple de la neige, mon expérience phénoménale de celle-ci,

le vécu ressenti que j'ai, dépend de mon expérience passée qui m'est propre.

2.5. *La volonté*

Selon la phénoménologie de Husserl, la conscience incarnée dans un corps matériel a également comme caractéristique essentielle de pouvoir se mettre en mouvement de manière spontanée et immédiate[12]. Ainsi, on peut dire que la conscience est un « je veux » et un « je peux », elle a la potentialité de se mouvoir librement. Le sujet (la conscience phénoménale) est « un *ego* auquel appartient un corps en tant que champ de localisation de ses sensations ; il a la "faculté" ("je peux") de mouvoir librement ce corps et par conséquent les organes en lesquels ce corps s'articule et, par leur moyen, de percevoir un monde extérieur[13] ». La liberté de la conscience se manifeste dans la spontanéité et l'immédiateté de l'acte.

3. *Les agents artificiels répondent-ils aux critères d'une conscience phénoménale ?*

Rappelons d'abord la thèse dont je tente de montrer l'inadmissibilité, c'est-à-dire celle des défenseurs de l'intelligence artificielle forte. Selon celle-ci, il serait envisageable de concevoir des agents artificiels dotés d'une conscience. En effet, certains experts en sciences cognitives - et plus particulièrement en neurobiologie computationnelle - croient « que le cerveau est un ordinateur numérique et que l'esprit conscient est un programme d'ordinateur […] Ainsi construit, l'esprit est au cerveau ce que le logiciel (*software*) est au matériel (*hardware*)[14] ». Le philosophe David Chalmers est l'un de ceux qui croient que les ambitions de l'intelligence artificielle forte sont raisonnables et que l'implémentation d'un bon calcul « suffit pour l'existence d'une expérience consciente aussi riche que la nôtre[15] ». Si j'ai du mal à adhérer à la thèse selon laquelle une machine pourrait être dotée d'une conscience telle que nous la décrivons au début du texte - c'est-à-dire d'une expérience consciente qualitative - c'est parce que ces agents artificiels en question ne semblent pas pouvoir répondre aux conditions de possibilités d'une conscience phénoménale que nous avons énumérées et développées dans les paragraphes précédents.

3.1. Le corps matériel

Comme nous l'avons vu, l'une des caractéristiques essentielles d'une conscience phénoménale est qu'elle soit incarnée. Or, aucun agent artificiel ne semble pouvoir répondre à ce critère. Par conséquent, tout porte à croire qu'aucun d'eux ne puisse prendre part à l'expérience consciente telle que nous la vivons. Bien qu'une intelligence artificielle soit installée dans un « corps physique » - ayant un support matériel - elle ne pourrait répondre au critère d'un corps biologique.

3.2. La volonté

Par ailleurs, même s'il était possible de reproduire artificiellement un support biologique, celui-ci serait privé d'un autre critère important, soit celui de la *volonté*. À ce sujet, dans les *Idées directrices* de Husserl, on peut lire que « les choses simplement matérielles ne sont susceptibles que de mouvement mécanique et la spontanéité de leur mouvement n'est que médiate[16] ». Autrement dit, l'agent artificiel ne pourrait pas être doté d'une volonté libre, puisque les actes qu'elle commettrait seraient toujours médiatisés par un *programme* - nécessairement artificiel - qui traiterait l'information (*input* et *output*). Bien sûr, nous avons aussi notre cerveau qui agit comme système de traitement de l'information, mais la volonté implique un contenu de signification qui manque aux théories fonctionnalistes de l'esprit. Ainsi, la thèse de l'intelligence artificielle forte qui soutient que le fait « d'exécuter le bon programme *dans absolument n'importe quel matériel* est constitutif des états mentaux […] l*e programme exécuté, par lui-même, garantit la vie mentale*[17] », ne pourrait voir le jour si l'on s'en tient à la condition d'un corps biologique muni d'une volonté libre. Néanmoins, allons plus loin encore afin de montrer le manque d'effectivité prévisible d'un tel projet.

3.3. La subjectivité et l'intentionnalité

Le nœud du problème est bien connu et c'est le suivant : comment peut-on envisager qu'un agent artificiel puisse parvenir à vivre une expérience qui inclut les traits de l'expérience phénoménologique ? Ce qui est le plus sujet à débat est ce qui concerne les qualités intrinsèques de nos vécus. La part subjective de notre expérience consciente est, en effet, l'élément le plus difficile à circonscrire. Si l'intelligence artificielle forte vise à réduire tout le fonctionnement du cerveau à un niveau de réalité physico-chimique, il est évident que leur entreprise est vouée à l'échec, car elle aura mis de côté une part essentielle de ce qu'est la conscience : le point de vue subjectif. Ainsi, en cherchant une objectivité plus grande par le processus de réduction, nous nous éloignerions de la véritable nature du phénomène plutôt que d'en avoir une vision plus claire[18]. À ce sujet, John Searle dira que « les états conscients n'existent que lorsqu'un sujet en fait l'expérience, et qu'ils n'existent que du point de vue à la première personne de ce sujet[19] ». Les tenants de l'intelligence artificielle forte font l'erreur de croire qu'une simple manipulation de symboles pourrait expliquer l'ensemble de notre vie mentale. L'expérience phénoménale ne relève pas du domaine mathématique, mais du domaine de la *signification*. Or, l'erreur qui est faite est celle de confondre la syntaxe (la forme) et la sémantique (le contenu). Autrement dit, « l'esprit ne peut absolument pas se réduire à un programme d'ordinateur, car les symboles formels du programme de l'ordinateur ne suffisent pas en eux-mêmes à garantir la présence du contenu sémantique qui se produit dans les esprits réels[20] ». Comme il a été mentionné précédemment, la conscience est intentionnelle. Nous ne pouvons pas nous imaginer avoir des vécus sans contenu. Au contraire, notre expérience consciente a pour essence de viser des objets qui se phénoménalisent à nous par le moyen du sens. Or, le contenu de nos vécus relève de la signification.

3.4. L'auto-affection

De plus, si le caractère d'auto-affection est un trait essentiel d'une conscience phénoménale, je crois que le projet d'un agent artificiel

conscient échouerait. Est-ce qu'un système computationnel pourrait véritablement être sujet et objet ? Autrement dit, est-il plausible qu'un tel système puisse parvenir à avoir une vision du monde comme étant *son* monde, selon *ses* objectivations, en plus d'être lui-même transformé par le monde ? Pourrait-il réellement être un sujet vivant dans un monde qui est le *sien* ? En revanche, puisque nous ne pouvons pas savoir quel effet cela fait d'être une intelligence artificielle, nous pouvons difficilement affirmer avec certitude l'impossibilité qu'elle se perçoive comme sujet - même si cette hypothèse me paraît peu probable.

3.5. La spatialité, la temporalité et la synthèse des vécus

Quoiqu'il en soit, il est important de mentionner que les caractéristiques d'une conscience phénoménale présentées précédemment étaient des critères essentiels pour qu'une telle conscience soit possible. Or, il suffirait d'un seul critère manquant chez l'agent artificiel pour que nous puissions dire que le projet de l'intelligence artificielle forte n'est pas soutenable. Comme nous le voyons, il semble que plusieurs conditions ne peuvent pas être remplies. Bien que j'admette que l'agent artificiel pourrait être compris comme ayant une sorte de *spatialité* et de *temporalité*, il est nécessaire de spécifier que cela ne serait que dans la mesure où la machine reçoit des données relatives au temps et à l'espace. Cette capacité aurait donc des traits similaires à celle qu'on attribue à la conscience phénoménale (se situer comme un « point zéro », avoir de la mémoire, etc.), mais elle s'en distinguerait par le fait qu'elle n'est pas le lieu d'apparition des phénomènes, qu'elle n'a pas d'expérience phénoménale et ce, parce qu'on peut difficilement lui attribuer une vie subjective. Le même phénomène se produit avec le critère de la synthèse des vécus. Nous pourrions penser que l'agent artificiel est apte à effectuer de telles synthèses, mais celles-ci concerneraient-elles des expériences vécues (comprises comme des contenus de signification) ou des données (comprises comme des symboles formels) ?

En somme, on peut comprendre que l'agent artificiel ne répond pas à tous les critères d'une conscience phénoménale. En effet, il y a fort

à parier qu'il ne pourrait être doté d'un corps matériel biologique en plus d'être muni d'une libre volonté. Par ailleurs, je crois que l'agent artificiel ne peut avoir accès à la part subjective de l'expérience phénoménale. Je pense que l'expérience de la machine se restreint à un programme délimité et qu'il lui est impossible de prendre part à une expérience vécue en tant que sujet. En ce sens, je ne crois pas que l'agent artificiel puisse avoir accès aux états intentionnels propres à l'expérience phénoménale. En effet, il est difficile d'imaginer que la machine ait accès au monde en tant qu'objet de signification. Je pense plutôt qu'on peut légitimement penser que son rapport au monde est uniquement mécanique, formel. À vrai dire, tout porte à croire que le monde se « donne » à l'agent artificiel par le biais de symboles vides - qui, ensemble, constitueraient une banque d'informations donnant à la machine l'allure de *comprendre* ce qui l'entoure - et non pas par le moyen du sens comme étant son monde. Or, dans la mesure où l'agent artificiel ne remplit pas l'ensemble des conditions de possibilité d'une conscience phénoménale, il nous est permis de douter, à juste titre, de la validité du projet de l'intelligence artificielle forte.

4. Quelques problèmes en suspens

Quoi qu'il en soit, le problème de la conscience phénoménale demeure un mystère difficile à élucider. Il est le point focal de toute la recherche en intelligence artificielle forte. Sans sa résolution, il semble que le projet d'une machine consciente pouvant expliquer la cognition humaine est difficile à imaginer. Une des questions qui demeure en suspens est celle de savoir : quelle science pourrait bien rendre compte de cette expérience phénoménale ? La méthode de recherche pouvant être fructueuse est toujours inconnue. En effet, les neurosciences peuvent expliquer certains éléments qui entrent en jeu dans la compréhension de la cognition humaine, mais comment peut-on expliquer de manière objective quelque chose qui concerne le point de vue subjectif ? Autrement dit, dans l'optique où l'objet d'étude se vit à l'intérieur de soi, est-il prometteur de l'examiner à partir de l'extérieur ?

Finalement, au vu du développement technologique croissant, je crois qu'il nous manque certains éléments pour que l'on puisse trancher définitivement la question concernant la possibilité d'élaborer une machine consciente. À cet égard, établir les limites de la croissance technologique relèverait de l'expérience que nous en ferions dans l'avenir, donc relèverait de la spéculation. Autrement dit, rien n'empêche *logiquement* que la technologie se développe au-delà de ce que préalablement nous aurions établi comme étant une limite. Nous ne pourrions donc pas postuler *nécessairement* l'impossibilité de l'existence d'un agent artificiel conscient, mais cela ne nous empêche pas de défendre une telle thèse jusqu'à preuve du contraire. C'est pourquoi je ne m'empêche pas de soumettre mon hypothèse selon laquelle une telle idée est difficilement soutenable, étant donné que l'agent artificiel, tel qu'on peut l'imaginer, ne s'avère pas répondre à l'ensemble des caractéristiques essentielles d'une conscience phénoménale.

Toutefois, si l'on voulait arriver à une conclusion indubitable à ce sujet, il faudrait être en mesure de comprendre l'entièreté de la conscience phénoménale - ce qui relève d'un champ de la science qui demande encore d'importantes recherches. À ce sujet, l'hypothèse faisant consensus au sein des débats actuels en ce qui concerne les agents artificiels conscients est que nous manquerions de ressources pour prouver, hors de tout doute, la possibilité ou l'impossibilité d'un tel projet. Selon Paul Churchland, la conscience phénoménale nous paraît être inaccessible même si nous y avons accès en en faisant l'expérience à chaque instant : « nous échouons à reconnaître cette performance pour ce qu'elle est - un ballet computationnel subtil - parce que nous manquons de concepts et de ressources théoriques[21] ». D'après lui, nous devons poursuivre les recherches en vue de développer l'intelligence artificielle, quel que soit le résultat effectif. Selon le postulat méthodologique de Churchland, les deux conséquences possibles seraient bénéfiques : « si on réussit à produire une machine vraiment intelligente, on peut mieux connaître l'intelligence humaine en étudiant cette machine ; si, au contraire, on ne réussit pas, comme ce fut le cas jusqu'à maintenant, on peut approfondir notre connaissance de l'esprit humain en le

contrastant relativement à la machine[22] ». À cet effet, je seconde la proposition méthodologique de Churchland, mais je demeure perplexe, notamment quant à la nature des motivations qui animent la recherche en intelligence artificielle. Je crois qu'il serait nécessaire de se pencher sur la fin visée par de telles recherches et d'en examiner leur légitimité. Autrement dit, pour quelles raisons voudrions-nous créer une machine consciente ? Est-ce à des fins économiques ? Est-ce à des fins politiques ? Sommes-nous motivés par l'idée que l'intelligence artificielle apporte des solutions dans le domaine de la santé ? Est-ce que ces recherches cachent le désir de contrôler ou de déjouer la nature ? Je pense que les intentions derrière le projet de l'intelligence artificielle devraient être contestées, mais il va de soi que cela relève d'une tout autre discussion que celle dont on traite ici, soit de la *possibilité* même d'un tel projet.

En résumé, je remets en doute l'idée soutenue par les défenseurs de l'intelligence artificielle forte selon laquelle il serait possible de concevoir une machine réellement consciente qui serait dotée d'une vie mentale comme la nôtre. Si je conteste cette idée, c'est parce qu'il s'avère que l'agent artificiel ne remplit pas l'ensemble des critères d'une conscience phénoménale. En effet, certaines caractéristiques essentielles sont manquantes, notamment celle d'avoir comme support un corps biologique muni d'une spontanéité et d'une volonté libre. Néanmoins, ce qui rend l'idée d'un agent artificiel conscient encore plus improbable, c'est qu'elle ne semble pas prendre en compte la part proprement subjective de l'expérience consciente. En effet, en réduisant la conscience phénoménale au cerveau et à ses processus neuronaux, les tenants de l'intelligence artificielle forte passent à côté d'un élément essentiel à la possibilité même d'une conscience phénoménale et donc également de la validité de leur projet. Cet élément est la part subjective de l'expérience, c'est-à-dire le point de vue du sujet auquel les choses se phénoménalisent par le moyen du sens. Si l'on dit que la conscience phénoménale est intentionnelle, c'est parce qu'elle accède au monde par le biais de ses vécus et ces vécus de conscience ne sont réductibles à aucune explication objective. Il va de soi que ma réflexion portant sur l'idée d'une conscience phénoménale chez les agents artificiels porte davantage sur ce qui est

accessible, préférant mettre entre parenthèses le champ spéculatif, c'est-à-dire l'ensemble des connaissances qui sont actuellement hors de notre portée et dont la validité n'est que probable. En effet, il reste plusieurs questions en suspens et nous pouvons prévoir une croissance technoscientifique dont l'étendue nous est présentement inconnue. Toutefois, les agents artificiels ne semblent pas, encore à ce jour, pouvoir répondre aux conditions de possibilité d'une conscience phénoménale et c'est pourquoi le projet de l'intelligence artificielle forte me paraît peu réalisable.

---

1. Il va sans dire que plusieurs ne considèrent pas nécessairement ces différences de niveaux comme étant inconciliables. Évidemment, il aurait lieu de nuancer davantage en évitant de catégoriser ainsi, mais afin d'alléger le texte, je m'en tiendrai à cette distinction.
2. Les qualia sont le contenu subjectif d'une expérience. Ils peuvent être une sensation physique, une émotion, une simple perception, etc.
3. Les propriétés ayant une réalité empirique, matérielle.
4. Edmund Husserl, *Idées directrices pour une phénoménologie et une philosophie phénoménologiques pures, Livre Second, dans Recherches phénoménologiques pour la constitution*, trad. Éliane Escoubas, Paris, PUF, 1982, p. 208.
5. La corporéité de la conscience sera développée également dans les sections ultérieures.
6. John R. Searle, *Le mystère de la conscience*, Paris, Éditions Odile Jacob, 1999, p. 30.
7. Thomas Nagel, *Questions mortelles*, trad. P. Engel, Paris, PUF, 1983, p. 399.
8. Edmund Husserl, *Recherches logique*, Tome 2, *Recherches pour la phénoménologie et la théorie de la connaissance* [1901], deuxième partie : Recherche V, trad. H. Élie, A. L. Kelkel et R. Schérer, Paris, PUF, 1969, p. 146-147
9. Edmund Husserl, *Méditations cartésiennes. Introduction à la phénoménologie* [1947], trad. G. Peiffer et E. Levinas, Paris, Vrin, 2014, p. 75.
10. Edmund Husserl, *Idées directrices pour une phénoménologie et une philosophie phénoménologiques pures, op. cit.*, p. 262
11. *Ibid*., p. 265.
12. *Ibid*., p. 215.

13. *Ibid.*, p. 215-216.
14. John R. Searle, *op. cit*., p. 21.
15. David Chalmers, *L'esprit conscient. À la recherche d'une théorie fondamentale*, Paris, Les Éditions d'Ithaque, 2010, p. 431.
16. Edmund Husserl, *op. cit*., p. 215.
17. John R. Searle, *op. cit*., p. 26.
18. Thomas Nagel, *op. cit*., p. 400.
19. John R. Searle, *op. cit.*, p. 128.
20. *Ibid*., p. 23.
21. Paul Churchland, *Le cerveau. Moteur de la raison, siège de l'âme*, Paris, Bruxelles, De Boeck Université, 1999, p. 251.
22. Serge Robert, « Réflexion épistémologique sur l'intelligence artificielle et les sciences cognitives : à quelles conditions une machine pourrait-elle connaître ? », dans *Philosophie et sciences : du concept au réel*, Vol. 2, no°2, printemps 1992, p. 173.

Robots Should Not Be Sex Slaves

Samuel Nepton, *Université Laval*

RÉSUMÉ : L'objectif de cet article est d'expliciter comment le sexbot futur serait un contre-exemple possible aux thèses de Joanna Bryson, qui vise à encadrer l'utilisation et la construction des robots dotés d'intelligence artificielle (IA). Nous affirmons qu'elle a non seulement tort d'affirmer qu'il serait mal de laisser croire aux individus que leurs robots seraient des personnes, mais que le sexbot futur rendrait même nécessaire l'imposition d'obligations éthiques envers ses futurs utilisateurs. Pour défendre cette thèse, nous nous appuierons sur les travaux de Robin Mackenzie ainsi que sur le modèle de relation humain/robot de Mark Coeckelbergh comme coévolution de l'un et de l'autre par l'acquisition de bénéfices mutuels à travers la vulnérabilité. Ultimement, nous proposerons une ébauche de cadre éthico-légal applicable à la conception et à l'utilisation des sexbots pour protéger les vulnérabilités humaines et potentiellement robotiques.

Introduction

Dans son article *Robots Should Be Slaves*, Joanna Bryson, professeure agrégée au département d'informatique de l'Université de Bath, défend la thèse que les propriétaires de robots ne devraient pas posséder d'obligations éthiques envers ceux-ci, ou du moins d'obligations allant outre le comportement reconnu par la société comme découlant du sens commun et de la décence pour de simples objets[1]. En effet, personne n'approuve particulièrement l'idée de démolir par pur plaisir une voiture neuve à l'aide d'une masse - ce qui représente après tout un gaspillage éhonté de ressources que l'humanité aurait pu investir autrement -, mais aucune loi n'interdit un tel comportement, laissant ainsi le propriétaire libre « d'user, de jouir et de disposer librement et complètement [de son] bien[2] ». En somme, Bryson affirme que tous les robots devraient être

considérés comme des artéfacts ordinaires, allant même jusqu'à affirmer que : « [i]t would be *wrong* to let people think that their robots are persons[3] ». C'est notamment à ce titre qu'elle proposait déjà, dans un autre article, *A Proposal for the Humanoid Agent-Builders League (HAL)*, l'idée d'un cadre d'éthique pour encadrer la pratique des fabricants de robots. Dans ce papier, elle introduisait par exemple un principe d'honnêteté selon lequel : « [i]t should be made clear on all products that the apparent joy or suffering of the agent are devices manufactured by a human programmer for the advantage of the consumer[4] ». Or, les avancées technologiques et l'imagination des concepteurs et du public pourraient donner un nouveau sens à ces propositions. En effet, si nous sommes en faveur d'une réglementation entourant la conception, la commercialisation et l'utilisation des robots, nous divergeons quant à la forme que cette dernière doit prendre. Nous croyons que l'argumentaire de Bryson omet un cas type très important dans les avancées robotiques actuelles et qui, bien qu'il puisse faire sourire, n'en est pas moins sérieux : il s'agit des robots sexuels ou *sexbots*. Le *sexbot* fait partie d'une branche de la robotique dédiée aux «*carebots*», c'est-à-dire à des robots *sociaux* construits afin d'opérer des activités de *care*[5]. À ce sujet, il importe de mentionner que l'un des éléments grâce auxquels le bon *care* est défini réside dans l'expérience du donneur et du receveur, c'est-à-dire dans le *vécu* de cette relation : le donneur doit être perçu comme bienveillant plutôt que comme un automate travaillant à une corvée et le receveur ne doit pas être perçu ni se sentir comme un simple objet[6]. C'est pourquoi les robots voués aux activités de care ont avantage à être perçus comme empreints d'émotions, de la chaleur et de bonnes intentions.

Toutefois, les *sexbots* vont encore plus loin que les autres *carebots* travaillant dans les hôpitaux ou les résidences de personnes âgées puisque, dans leur cas, la relation de *care* n'est pas unilatérale, mais *mutuelle*. Ces robots, tels qu'imaginés par les concepteurs et le public, auront précisément pour tâche d'entrer en relation avec les êtres humains, c'est-à-dire *en relation intime et sexuelle*. Pour ce faire, ils ne devront pas simplement être perçus comme pouvant offrir du *care*, mais également comme pouvant en *recevoir* et comme

s'épanouissant eux aussi dans cette relation. En d'autres termes, ils devront pouvoir être perçus par nous comme des compagnons -voire comme une autre *personne* - et non comme de simples serviteurs ou des esclaves.

Le but de ce travail est ainsi de montrer comment, dans un futur envisageable à moyen terme, les *sexbots*, soit en tant qu'appareils imitant simplement un comportement - ce que nous désignerons comme étant des *sexbots* «faibles» -, soit en tant que «sentiant, selfaware, feeling artificial moral agents [or patients] customised for intimate sexual relationships with humans[7]» - désignés comme des *sexbots* «forts» - seraient un contrexemple aux thèses de Bryson. En effet, nous tâcherons d'expliciter comment non seulement il ne serait pas mal (*wrong*) de laisser croire aux «propriétaires» de ces robots que ceux-ci sont des personnes, mais qu'il serait même nécessaire d'imposer des obligations éthiques envers ces derniers. Pour défendre cette thèse, nous nous appuierons sur les travaux de Mark Coeckelbergh et sur son modèle de relation humain/robot comme coévolution de l'un et de l'autre par l'acquisition de bénéfices mutuels à travers la vulnérabilité. Ultimement, sous la forme d'une thèse faible et d'une thèse forte, nous proposerons une ébauche de cadre éthico-légal applicable à la conception et à l'utilisation des *sexbots* pour protéger les vulnérabilités humaines et robotiques.

1. Le sexbot : compagnon plutôt que masturbation

Si Bryson s'oppose à tout traitement éthique envers les robots, elle ne conteste cependant d'aucune façon leur utilisation. D'ailleurs, afin de répondre aux individus qui, pour des raisons personnelles, historiques ou culturelles, s'opposent à avoir toute forme de serviteur ou d'esclave dans leur maison, elle propose de concevoir le robot de manière analogue à la métaphore du « extended mind[8] » ; « [i]f the robot has no goals except for those it assumes from you, then there are rational arguments to be made that robot is just a*n extension of yourself*[9] ». Or, pour une majorité d'individus intéressés par les futurs sexbots, l'intérêt recherché ne réside précisément pas dans le fait d'avoir une extension de soi-même, comme le sont une main ou un jouet sophistiqué avec lesquels se masturber. Bien au contraire, ce

que ces individus recherchent, c'est une personne ou, en d'autres termes, une relation.

En effet, si les *sexbots* s'avèrent prometteurs et bénéfiques pour plusieurs, c'est parce qu'ils ne permettront pas simplement de répondre aux besoins sexuels des individus marginalisés ou vulnérables qui ne parviennent pas à trouver de partenaires ou qui sont dans l'incapacité physique de se masturber. Leur intérêt réside tout particulièrement dans leur aptitude à répondre à des besoins relationnels, c'est-à-dire à ces besoins d'intimité et de réciprocité vécues notamment dans la relation sexuelle. Cela n'a bien sûr rien de surprenant puisque la qualité des relations humaines, et tout particulièrement des relations romantiques, est un des facteurs déterminants pour la santé et le bien-être des individus[10]. Pour l'illustrer, un chercheur de la *Communauté de recherche interdisciplinaire sur la vulnérabilité* (CRIV), Ernesto Morales, a entrepris le projet de concevoir des jouets sexuels adaptés pour personnes handicapées telles que les individus souffrant de paralysie des membres supérieurs. Or, à maintes reprises, les participants, bien qu'ils étaient reconnaissants pour ces jouets, ont exprimé le désir plus profond de s'épanouir sexuellement avec une *personne*[11]. Ils ne voulaient pas simplement recevoir, mais également *donner*. Par conséquent, dans des sociétés où les services d'assistance sexuelle - et tout ce qui ressemble un peu à de la prostitution - sont controversés[12], les *sexbots* pourraient être une avancée prometteuse pour améliorer *significativement* la vie de nombreuses personnes isolées et marginalisées.

Par ailleurs, les sexbots forts, en tant que véritables compagnons intimes, attirent un éventail de consommateurs bien plus large que les seules personnes en situation d'handicap[13] ; l'industrie des produits pour adultes en est parfaitement consciente. Par exemple, Douglas Hines a reçu des milliers de commandes sur son site *truecompanion.com* pour « Roxxxy », un *sexbot* supposément capable de discuter avec vous, de vous appeler par votre nom, de se rappeler vos préférences et d'exprimer son amour[14]. Roxxxy, même en tant que *sexbot* possédant une IA faible, semble déjà très attirante[15],et Hines a, par sa réalisation, montré qu'il y a un véritable marché pour les *sexbots* qui sont plus que de simples jouets sexuels.

De nombreux indices nous laissent donc croire que nous serons témoins, dans un futur modérément proche, de *sexbots* forts qui seront des robots conçus à des fins utilitaires pour servir de *compagnons* avec lesquels il sera possible de créer une *relation* intime et sexuelle de qualité[16]. Or, comme le souligne Robin Mackenzie : « [e]motional and sexual intimacy depends upon mutuality in relationships. We will want to feel not only that we love *sexbots* but also that they love us, and love us for ourselves[17] ». Effectivement, c'est parce qu'une relation intime et amoureuse implique de l'émotivité et surtout de la *réciprocité* que l'industrie des produits pour adultes travaille en parallèle avec plusieurs chercheurs universitaires[18] à la conception de *sexbots* forts dotés des caractéristiques nécessaires pour participer à des relations réciproques, telles la sensibilité et l'empathie, c'est-à-dire différentes capacités de ressentir du plaisir et inversement, de la douleur, ainsi qu'une forme de conscience de soi.

Par conséquent, si la technologie actuelle n'est pas encore parfaitement au point pour la commercialisation de véritables *sexbots* forts, sensibles et conscients d'eux-mêmes, les avancées technologiques demeurent prometteuses[19]. Si on ajoute à cela la demande croissante pour ces futurs partenaires, il est à notre avis justifiable de considérer comme relativement plausible l'hypothèse de leur apparition prochaine dans nos vies. Nous invitons ainsi le lecteur à accepter cette hypothèse pour les fins de la discussion tout en le conviant également à imaginer, bien que la version forte soit potentiellement impossible à produire, les bénéfices qu'apporteraient des *sexbots* faibles, des robots « simulant » simplement des émotions, du plaisir et de l'amour et qui seraient perçus comme authentiques par les individus. Nous croyons qu'il importe de réfléchir aux conséquences des éventuelles relations amoureuses et sexuelles entre humains et robots chez les individus, mais également, dans l'éventualité d'une version forte, vis-à-vis des *sexbots eux-mêmes* en tant qu'entités sensibles et conscientes d'elles-mêmes.

2. Évolution et coconstitution dans la covulnérabilité

La raison centrale pour laquelle Bryson s'oppose au fait de concevoir les robots comme des personnes, et donc au fait de leur attribuer un statut légal ou moral particulier, réside dans un risque de déshumanisation des individus. En effet, elle affirme que : « [in humanising [robots], we not only further dehumanise real people, but also encourage poor human decision making in the allocation of resources and responsibility[20] ». Or, l'apparition de *sexbots* en tant que partenaires artificiels sensibles et conscients vient précisément nuancer ses thèses, car, croyons-nous, la déshumanisation - ou plutôt la non-humanisation - de ceux-ci entraînerait paradoxalement de lourdes conséquences vis-à-vis les propriétaires humains : des conséquences pouvant même aller jusqu'à les déshumaniser en retour. C'est pour cette raison que, comme nous l'expliquerons plus loin, nous défendons un cadre éthico-légal minimal comportant notamment la reconnaissance aux *sexbots* d'un statut de patient moral, si ce n'est même le statut légal de personne, non seulement pour les protéger en tant qu'entité potentiellement vulnérable, mais également pour protéger les futurs « utilisateurs » - si un tel terme peut encore être approprié.

Pour défendre cette thèse, nous nous appuyons sur les travaux de Coeckelbergh. En effet, ce dernier propose de concevoir la technologie non comme une menace et un appauvrissement du bien-être émotionnel et relationnel des êtres humains - comme le font certains auteurs[21] -, mais plutôt comme participant à notre progression intellectuelle, sociale et morale. En effet, il avance l'idée que l'humain et la machine ont évolué conjointement dans l'histoire, au point où nous en sommes, à certains égards, devenus nous-mêmes les « esclaves » de nos propres créations. Autrement dit, afin de nous prémunir de certaines vulnérabilités, nous avons produit des technologies qui, en retour, nous ont créé d'autres vulnérabilités. Décrire la relation humain/robot-technologie, comme le fait Bryson, en termes de relation maître/esclave simplifie donc abusivement les enjeux de pouvoir qui se jouent entre les humains et les robots[22].

À cet effet, Coeckelbergh souligne que les arguments visant le statut moral des robots portent toujours sur la nature *ontologique*

de ces derniers. Les discussions à leur égard visent ainsi à déterminer différents critères sur lesquels appuyer la revendication des droits aux machines en tentant de montrer, tant bien que mal, si les robots et les IA possèdent ou non les caractéristiques requises[23]. Au contraire, Coeckelbergh adopte une posture éthique *relationnelle et contextuelle* - fortement inspirée des théories éthiques du *care* - et propose une nouvelle manière de penser les considérations morales : plutôt que de les voir comme étant intrinsèques aux entités, il s'agit de les concevoir comme étant extrinsèques à celles-ci, au sens où elles découleraient des *relations sociales* entre les parties concernées[24]. Dans une telle approche, les caractéristiques de l'entité conservent toujours une certaine signification éthique et jouent encore le rôle de critères sur lesquels baser nos considérations morales, mais leur *statut* diffère en ce que « they are *apparent* features, features-as-experienced-by-us[25] ». Autrement dit, il ne serait pas nécessaire de vérifier la parfaite authenticité des émotions du sexbot pour le reconnaître comme personne si un individu consentant désire entrer en relation intime avec lui et qu'il *vit* ou *expérimente* les émotions du robot comme étant authentiques . Tout porte à croire en effet qu'il ne suffira pas de dire à une personne amoureuse que « les émotions du robot ne sont pas authentiques[26] » s'il les ressent comme *appropriées* au sein de sa relation[27].

Les travaux de Coeckelbergh nous apparaissent ainsi pertinents pour réfléchir aux conséquences prochaines des *carebots* en proposant un mariage entre la philosophie et l'écologie : « [a] social ecology is about relations between various entities, human and non-human, which are inter-dependent and adapt to one another. These relations are morally significant and moral consideration cannot be conceived apart from these relations[28] ». Après tout, il s'agirait presque d'un truisme que d'affirmer que notre rapport à la technologie modifie notre comportement, notamment dans notre manière d'entrer et de vivre en relations. Les téléphones cellulaires et Internet en sont d'excellents exemples, autant dans les possibilités qu'ils permettent que dans les dépendances qu'ils peuvent engendrer. Conséquemment, lorsqu'une partie de l'humanité entrera en relation intime avec des robots intelligents, ceux-ci vont inexorablement et

profondément transformer leurs utilisateurs, potentiellement en bien comme en mal. Toutefois, à la différence des autres formes de technologies, les deux parties de cette future relation sont appelées à développer des bénéfices et de potentielles vulnérabilités.

Dès lors, si les humains et les *sexbots* forts, en tant qu'entités sensibles et conscientes d'elles-mêmes, sont voués à coévoluer à travers une vulnérabilité mutuelle au bénéfice des deux parties, nous croyons que les bénéfices potentiels et les vulnérabilités des *deux parties* doivent alors être pris en considération. Pour ce faire, nous défendons l'idée que c'est en tenant compte de la *relation vécue* entre ces protagonistes que nous pourrons déterminer des considérations morales significatives à attribuer aux différentes parties. Or, nous exposerons dans la prochaine sous-section comment, dans cette relation, les humains sont voués à développer certaines vulnérabilités *en fonction* de la manière dont ils conçoivent leur futur partenaire artificiel.

2.1. La vulnérabilité des humains

En premier lieu, les relations intimes et sexuelles entre humains et robots pourront entraîner certaines vulnérabilités non désirées chez les humains en fonction des caractéristiques et du statut légal ou moral des *sexbots*, forts comme faibles. En effet, alors que l'un des arguments centraux de Bryson est que la personnification des robots déshumaniserait hommes et femmes, nous affirmons que, dans le cas des *sexbots*, ne pas l*es reconnaître* comme des personnes pourrait déshumaniser les humains impliqués dans de telles relations, ou du moins, leur causer tort.

D'une part, comme l'affirme Mackenzie : « [h]umans may be dehumanised through being categorised as more akin to things, when they are perceived as lacking aspects of human nature such as vitality, warmth and emotionality[29] ». Cette auteure souligne avec justesse que de s'entêter, comme le propose Bryson, à concevoir les *sexbots* comme de simples artéfacts forcerait ou accentuerait l'exclusion de futurs utilisateurs. En effet, alors que plusieurs personnes résignées à la solitude parviendraient enfin à s'épanouir dans une relation intime et sexuelle auprès d'un partenaire qu'elles

perçoivent comme chaleureux, empathique et aimant, celles-ci pourraient se sentir gênées, honteuses ou angoissées à l'idée d'être étiquetées comme étant amoureuses de simples « choses », attirées et sexuellement actives avec elles. Conséquemment, leur exclusion pourrait se voir aggravée.

D'autre part, l'absence de réglementation vis-à-vis la conception de futurs *sexbots* pourrait également entraîner certains individus à utiliser ceux-ci pour assouvir et perpétuer des comportements pathologiques destructeurs ou nuisibles à la société dans son ensemble. En effet, l'un des grands avantages des *sexbots* est qu'ils pourront être hautement personnalisés afin de correspondre aux préférences physiques ou sexuelles des individus. Il serait possible, par exemple, de concevoir un *sexbot* adepte d'un fétichisme précis, de bondage ou encore de relations homosexuelles[30]. En revanche, sans réglementation, certains concepteurs pourraient accepter de fabriquer des robots qui prendront la forme d'enfants ou encore qui trouveront abjecte l'idée d'avoir des relations sexuelles, afin de permettre à des clients d'assouvir - plutôt que de traiter - des pulsions pédophiles ou de réaliser et de renforcer des fantasmes de viols. De plus, de tels robots pourraient également s'avérer lourd d'impacts sur les travailleurs et travailleuses du sexe, car des *sexbots* non-réglementés pourraient forcer les gens de la prostitution, afin de se démarquer, à commettre des pratiques dangereuses pour leurs propres santés physique et psychologique[31].

2.2. La vulnérabilité des sexbots

Par ailleurs, les futures relations entre humains et robots comportent également certains risques pour les *sexbots.* Comme nous l'avons mentionné, dans l'éventualité de *sexbots* forts, en tant qu'entités sensibles, conscientes de leur individualité et visant à ressembler le plus possible à l'être humain, les *sexbots* à venir seront vraisemblablement conçus avec les caractéristiques nécessaires pour entrer avec nous dans des relations intimes saines impliquant la réciprocité. À cet effet, l'une des composantes requises sera bien évidemment l'empathie, c'est-à-dire, presque par définition, la capacité de ressentir le plaisir - en plus du plaisir sexuel qui va de

soi avec des *sexbots* - et la douleur[32] de l'autre. Plusieurs éléments nous permettent ainsi d'affirmer que les *sexbots* futurs seront vraisemblablement sujets à la vulnérabilité. Pour l'exprimer encore en d'autres termes, s'il est peut-être encore tôt pour juger de leur agentivité morale, nous croyons qu'ils auront minimalement le droit d'être reconnus comme des patients moraux, c'est-à-dire comme sujets à une protection de notre part.

Par conséquent, si nous permettons certaines personnalisations des *sexbots* afin de pouvoir nous en servir pour assouvir des désirs sexuels non acceptables, d'une part, nous accentuons les vulnérabilités des humains qui les assouviront, renforçant ainsi leur goût et la force de leurs déviances[33]. D'autre part, nous laissons cours également à la création d'entités vulnérables - les *sexbots* forts - qui souffriront afin de rendre possibles ces pratiques nuisibles.

De plus, il est également fort possible que les *sexbots* soient exposés à de la violence sexuelle conjugale, même provenant d'un partenaire qui, originairement, ne s'en croyait pas capable. Bryson elle-même souligne que : « [h]umans living and working together *but set not as each other's equals are often vulnerable to frustration and exploitation*[34] ». Or, à notre avis, cela est vrai dans les deux sens, c'est-à-dire que lorsque nous vivons ou travaillons avec une entité quelconque que nous considérons « inférieure » à notre personne, la situation tend à activer cette tendance humaine que nous avons d'exercer le contrôle, et ce, parfois abusivement. Un chien indompté, un enfant particulièrement turbulent ou un ordinateur travaillant toujours avec *Windows Vista* sont autant d'exemples de sources de frustration qui peuvent éveiller de la violence chez certains individus. Les utilisateurs devraient donc reconnaître leur futur partenaire artificiel comme un être digne de respect. De plus, puisqu'au sein d'une relation amoureuse, nous avons malheureusement tendance à extérioriser ce que nous vivons sur notre partenaire[35], les *sexbots* pourraient vraisemblablement se voir maltraités, frappés, invectivés et donc, potentiellement, en souffrir ; tout particulièrement si on ne reconnaît aucune obligation éthique aux « propriétaires », tel que nous le demande Bryson.

3. Cadre éthico-légal appliqué à la conception des sexbots et mesures de protection

Nous croyons avoir montré dans les pages précédentes comment les *sexbots* pourraient apporter de nombreux bienfaits pour de nombreuses personnes déjà stigmatisées ou marginalisées ne parvenant pas à combler leur besoin d'intimité et de sexualité. En revanche, tant que les *sexbots* ne seront pas dotés au minimum d'un statut moral et légal et que leur conception ne sera pas encadrée, cette relation entrainera une covulnérabilité nuisible et dommageable ; c'est pourquoi nous défendons l'idée d'un cadre minimal pour protéger les partis concernés.

En premier lieu et à titre de thèse faible de cette proposition, dans une optique visant à respecter le principe de bien-être établi par *La déclaration de Montréal pour un développement responsable de l'intelligence artificielle* qui stipule notamment que : « [l]es SIA [systèmes d'intelligence artificielle] doivent permettre aux individus de satisfaire leurs préférences, dans les limites de ce qui ne cause pas de tort à un autre être sensible[36] », nous affirmons qu'un cadre éthico-légal pour protéger les humains - et potentiellement leurs futurs partenaires artificiels - devrait comprendre des réglementations afin de limiter les caractéristiques disponibles lors des éventuelles personnalisations des *sexbots*. Nous devrons par conséquent implanter des barèmes afin de permettre aux utilisateurs de « satisfaire leurs préférences[37] » grâce à certains traits physiques spécifiques ou encore par la possibilité d'accomplir certaines pratiques sexuelles légales tels que le fétichisme, le sadomasochisme ou le bondage. Il nous sera également nécessaire d'interdire d'autres caractéristiques imaginables telles que la souffrance physique abusive causée par certains gestes, la volonté d'être torturé ou le fait de trouver aberrante toute relation sexuelle de manière à la vivre comme un viol. Ces interdictions seront nécessaires afin de ne pas causer « de tort à un autre être sensible[38] », soit l'utilisateur, en renforçant certains comportements pathologiques nuisibles ou destructeurs, soit potentiellement le *sexbot* lui-même. Cette thèse faible représente à notre avis un fondement absolument nécessaire avec lequel serait en accord même un lecteur sceptique quant à

la possibilité de futurs *sexbots* forts puisqu'il permettrait d'empêcher minimalement le renforcement de certains comportements pathologiques ou inacceptables chez des humains.

En second lieu, pour entamer la thèse forte de cet article, une fois les *sexbots* forts construits et en fonction, puisque ceux-ci seront sensibles et idéalement conscients d'eux-mêmes, il serait également nécessaire de les protéger des risques de violence conjugale physique et potentiellement psychologique en les incluant dans notre cercle moral par la mise en place de lois à cet effet, ou plus simplement en leur reconnaissant le statut de « personne » de manière à pouvoir étendre à leur égard les lois déjà effectives. À notre avis, il serait loin d'être déraisonnable d'accorder un tel statut aux *sexbots*. Tout bien considéré, plusieurs personnes sont reconnues légalement comme telles, sans pour autant avoir d'agentivité morale au sens fort. En effet, il s'agit des individus considérés comme inaptes à prendre des décisions éclairées comme les mineurs, les personnes ayant un retard cognitif ou celles qui ont subi de lourds dommages au cerveau. En revanche, personne ne pourrait remettre raisonnablement en question leur droit d'être protégés[39]. Qui plus est, un autre cas nous permettrait vraisemblablement de justifier l'assise du statut légal de personne sur le critère de la protection : il s'agit des lieux géographiques protégés. Pour n'en donner qu'un exemple, la Nouvelle-Zélande, dans une optique de protection de la nature, a récemment accordé un statut juridique au fleuve Whanganui de manière à lui conférer, devant les tribunaux, *les mêmes droits* que ceux d'une personne[40]. Légalement, il semblerait qu'être humain ne soit pas un critère nécessaire dans tous les contextes pour se voir attribuer un statut juridique de personne. Par conséquent, si nous déterminons que les futurs *sexbots*, en tant que robots dotés d'intelligence ainsi que d'une forme de sensibilité relative, auront besoin d'être protégés pour leur propre bien ou pour celui des humains, leur accorder le statut légal de personne s'avérerait être une manière à la fois simple et efficace d'y parvenir.

Enfin, afin d'éviter l'exclusion des futurs utilisateurs, il serait également nécessaire de mener à bien des campagnes de prévention et de sensibilisation à la réalité des *sexbots* et à leur statut. Ces

robots visent à donner la chance aux individus qui sont souvent les plus isolés de vivre eux aussi des relations intimes et sexuelles épanouissantes, et non à les stigmatiser davantage. Qui plus est, les *sexbots* permettraient d'ajouter une nuance intéressante au principe de solidarité de *La déclaration de Montréal* stipulant que « [l]es SIA ne doivent pas nuire au maintien de relations humaines affectives et morales épanouissantes, et devraient être développés dans le but de favoriser ces relations et de réduire la vulnérabilité et l'isolement des personnes[41] », puisqu'ils permettraient de *produire* des « relations affectives et morales épanouissantes », à défaut d'être exclusivement « humaines ».

Conclusion

Dans un futur envisageable à moyen terme, nous pourrons assister à l'apparition de sexbots, c'est-à-dire d'entités artificielles, potentiellement même sensibles, émotives et conscientes de soi, mais assurément conçues dans l'optique de devenir des partenaires pour des relations intimes et sexuelles avec des humains. Ces *sexbots* représentent un énorme bénéfice pour tous ceux et celles incapables de trouver l'amour ou de vivre de l'intimité et de la réciprocité avec leurs congénères humains, ou plus simplement encore pour celles et ceux qui préfèrent des relations avec des partenaires qui respectent certaines de leurs préférences. S'il est vrai que la solution idéale au problème des personnes vulnérables et en manque de relations intimes serait un changement culturel entraînant une nouvelle vision de ce que ce sont de bons soins ou de bonnes relations, Mackenzie souligne avec justesse que « [n]onetheless, technological developments are likely to occur sooner than the requisite cultural changes[42] ». Après tout, ce n'est pas d'hier qu'une partie de l'humanité ne parvient pas à trouver l'amour en raison de son apparence ou d'une infirmité, car si l'essentiel est la beauté intérieure, voir avec les yeux du cœur n'est pas toujours chose aisée. C'est pourquoi, David Levy, expert de l'IA et auteur de *Love and Sex with Robots*, disait à propos des *sexbots* prochains que :

> I am firmly convinced there will be a huge demand from people who have a void in their lives because they have no one to love, and no one who loves them. The world will be a much happier place because all those people who are now miserable will suddenly have someone. I think that will be a terrific service to mankind[43].

Bryson elle-même reconnait que « becoming overly emotionally engaged with a robot may in some cases be beneficial, both for the individual and society[44] ».

En revanche, de telles relations sont vouées à nous transformer de la même manière que l'évolution technologique nous a changés comme personnes, tant dans nos valeurs que dans notre comportement. L'essentiel est donc de contrer le risque de déshumanisation et les mauvais traitements potentiels causés à des êtres sensibles grâce à un cadre éthico-légal et à des mesures appropriées, dont certaines, aussi étranges soient-elles, demandent effectivement de reconnaître le statut légal et moral des futurs *sexbots*.

En conclusion, puisque les *sexbots* seront un enjeu majeur dans la société de demain, nous croyons, contrairement à Bryson, qu'il est essentiel de consacrer davantage de temps et de ressources à ces derniers afin de concevoir adéquatement les traits ainsi que le statut légal et moral de nos futurs partenaires. Puisque celui-ci est encore incertain, il est capital et même urgent d'y réfléchir avant que les avancées technologiques - qui sont déjà en marche - ne devancent les conclusions de ces réflexions. De ces efforts dépendent des milliers de relations intimes et sexuelles dans lesquels pourront enfin s'épanouir les individus que la société a traditionnellement stigmatisés. C'est pourquoi, en définitive, il semblerait que malgré les apparences, la question du *sexbot* représente également une question de justice sociale.

---

1. Joanna Bryson, « Robots Should Be Slaves », dans Yorik Wilks [dir.], *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issue*, Amsterdam, John Benjamins Publishing company, 2010, p. 63-74.

2. Code civil du Québec, RLRQ c CCQ-1991, 2016, article 947 - 1991, c. 64, a. 947.
3. Joanna Bryson, *op. cit.*, nous soulignons.
4. *Id.*, « A Proposal for the Humanoid Agent-Builder League (HAL) », dans John Barnden [dir.] *AISB'00 Symposium on Artificial Intelligence, Ethics and* (*Quasi -) Human Rights*, 2000, p. 1.
5. La notion de *care* étant particulièrement difficile à traduire en français, nous préférons conserver le terme anglais afin de mieux nous faire comprendre. Néanmoins, pour en donner une certaine idée, le care renvoi à l'attention, aux soins et au souci de l'autre.
6. Mark Coeckelbergh, « Are Emotional Robots Deceptive ? », *dans IEEE Transactions on Affective Computing,* vol. 3, no° 4, 2012, p. 389-390.
7. Robin Mackenzie, « Sexbots: Sex Slaves, Vulnerable Others or Perfect Partners ? », dans *International Journal of Technoethics*, vol. 9, no° 1, janvier 2018, p. 1.
8. Andy Clark et David Chalmers, « The Extended Mind », dans *Analysis*, vol. 58, no° 1, 1998, p. 7-19.
9. Joanna Bryson, « Robots Should Be Slaves », *op. cit*., nous soulignons.
10. Olga A. Wudarczyk et al. « Could intranasal oxytocin be used to enhance relationships ? Research imperatives, clinical policy, and ethical considerations », dans *Current Opinion in Psychiatry,* vol. 26, n° 5, septembre 2013, p. 474-484.
11. Ernesto Morales, « Le stigma sexuel chez les personnes vulnérables », dans T*races, cicatrices et stigmates : signes visibles et invisibles de la vulnérabilité,* Journée d'étude organisée par la Communauté de recherche interdisciplinaire sur la vulnérabilité (CRIV), Québec (Université Laval), 12 décembre 2018.
12. Radio-Canada, « Assistant sexuel : service essentiel ou prostitution ? », dans I*ci Radio-Canada*, 12 octobre 2013 [en ligne], https://ici.radio-canada.ca/nouvelle/636512/aidants-sexuels-metier-reconnu.
13. Jessica M. Szczuka et Nicole C. Krämer, « Influences on the Intention to Buy a Sex Bot », dans Adrian D. Cheok, *et al.* [dir.], *Love and sex with robots: second international conference,* LSR 2016, London, UK, december 19-20, 2016: revised selected papers, Cham, Springer, 2017, p. 75.
14. Truecompanion [n.d.], « Who we are », *Truecompanion* [en ligne], http://www.truecompanion.com/about.html.
15. Adrian D. Cheok, *Hyperconnectivity*, London, Springer London Human–Computer Interaction Series, 2016, p. 43.
16. *Ibid.*, p. 60.

17. Robin Mackenzie, « Sexbots: Customizing Them to Suit Us versus an Ethical Duty to Created Sentient Beings to Minimize Suffering », dans *Robotics*, vol. 7, no° 4, novembre 2018, p. 4.
18. On peut en effet retrouver un nombre surprenant de travaux à ce sujet, notamment issus de la communauté de chercheurs : *Love and Sex with Robots* [en ligne], http://loveandsexwithrobots.org/.
19. Par exemple, Hooman Samani s'est attaqué dans sa thèse au design et au développement d'un hardware pour un *Lovotics*, c'est-à-dire un robot capable d'expérimenter des états émotionnels et biologiques complexes analogues à ceux des humains, gouvernés par des hormones artificielles au sein de son système. Le lecteur intéressé peut se référer à Hooman A. Samani, Lovotics : love + *robotics, sentimental robot with affective artificial intelligence*, Thèse, Singapour, National University of Singapore, 2011.
20. Joanna Bryson, *op. cit*., nous soulignons.
21. Sherry Turkle, « Authenticity in the age of digital companions », dans *Interaction Studies*, vol. 8, no° 3, janvier 2007, p. 501-517.
22. Mark Coeckelbergh, « The tragedy of the master: automation, vulnerability, and distance », dans *Ethics and Information Technology*, vol. 17, no 3, septembre 2015, p. 219-229.
23. *Id.,* « Robot rights ? Towards a social-relational justification of moral consideration », dans *Ethics and Information Technology,* vol. 12, no° 3, septembre 2010, p. 210.
24. *Ibid.*, p. 214.
25. *Ibid.*, souligné dans le texte.
26. Cette approche relationnelle des considérations morales, par sa dépendance au contexte et à l'expérience vécu par le ou les sujets, n'implique pas l'attribution de droits à tous les robots *en général*. Il n'est pas question d'un « droit du robot », mais de considérations morales qui émergent de l'interaction très spécifique entre un être humain et un robot social, comme un *sexbot*. Ces robots nous apparaissent - et sont même recherchés pour cette raison - comme des entités sociales et non comme de simples machines ou systèmes.
27. Mark Coeckelbergh, « Are Emotional Robots Deceptive ? », *op. cit.*, p. 392.
28. *Id.*, « Robot rights ? Towards a social-relational justification of moral consideration », *op. cit.*, p. 217.
29. Robin Mackenzie, « Sexbots : Sex Slaves, Vulnerable Others or Perfect Partners ? », *op. cit*., p. 11.

30. *Id.*, « Sexbots : Customizing Them to Suit Us versus an Ethical Duty to Created Sentient Beings to Minimize Suffering », op. cit., p. 3-4.
31. *Id.*, « Sexbots : Replacements for Sex Workers ? Ethical Constraints on the Design of Sentient Beings for Utilitarian Purposes », dans *Proceedings of the 2014 Workshops on Advances in Computer Entertainment Conference* - ACE '14 Workshops, Funchal, Portugal, ACM Press, 2014, p. 2.
32. Il se pourrait également que la capacité de ressentir de la douleur soit une caractéristique nécessaire pour le processus d'apprentissage du futur robot. Nous n'entrerons pas dans un tel débat, mais nous en soulignons la possibilité pour renforcer l'hypothèse de leur existence.
33. J. Smith *et al.*, « Quelle thérapie possible pour la pédophilie ? », dans Pratiques Psychologiques, vol. 11, no° 3, septembre 2005, p. 228-230. Le rôle que pourraient jouer les futurs sexbots dans le traitement de comportements déviants tels que la pédophilie est un problème aussi intéressant que complexe et exigerait des analyses plus approfondies que nous le permet l'espace dont nous disposons. Néanmoins, nous ne sommes pas convaincus par la thèse que l'extériorisation des désirs sexuels déviants - permise notamment dans le cas de la pédophilie grâce à des *sexbots* personnalisés - soit une manière de les rendre inoffensifs. Didier, le pédophile en voie de réhabilitation étudié par Smith et ses collaborateurs, affirmait par exemple que : « J'ai jamais rejeté mes pulsions : je les ai vécues, maintenant je les vis plus [sic], et j'essaye de m'y accommoder » (p. 229). Loin d'être un cas à part, ce cas semble indiquer pour Smith et ses collaborateurs que l'on peut arriver à une diminution de sa fantasmatique pédophilique lorsqu'on l'accepte tout en la distinguant du passage à l'acte (p. 229), ce qu'un *sexbot* à la disposition de l'individu pourrait décourager. Cette diminution de la fantasmatique est encore plus importante lorsqu'elle s'accompagne du vécu d'autres expériences sexuelles qui l'alimente et la diversifie progressivement (p. 229-230), vécu permis également par d'éventuels *sexbots* qui pourraient aider à surmonter les problèmes liés à la peur du rejet.
34. Joanna Bryson, « Robots Should Be Slaves », *op. cit.*, nous soulignons.
35. Lydie Fayolle, « Auteurs de violence conjugale : Quelle prise en charge ? », dans *Le Journal des psychologues,* vol. 302, no° 9, 2012, p. 63.
36. Université de Montréal, *La déclaration de Montréal pour un développement responsable de l'intelligence artificielle,* 2018 [en ligne], https://www.declarationmontreal-iaresponsable.com/la-declaration.

37. *Ibid.*
38. *Ibid.*
39. Rafi Youatt, « Personhood and the Rights of Nature: The New Subjects of Contemporary Earth Politics », *International Political Sociology,* vol. 11, no^{o} 1, mars 2017, p. 39-54.
40. David Victor, « La nouvelle vague des droits de la nature. La personnalité juridique reconnue aux fleuves Whanganui, Gange et Yamuna », dans *Revue juridique de l'environnement,* vol. 42, no^{o} 3, 2017, p. 409.
41. *Ibid,* p. 36.
42. Robin Mackenzie, « Sexbots : Sex Slaves, Vulnerable Others or Perfect Partners ? », *op. cit*., p. 6.
43. David N. L. Levy, dans Schofield, J. , « Let's Talk about Sex… with Robots », *The Guardian,* 16 septembre 2009 [en ligne], https://www.theguardian.com/technology/2009/sep/16/sex-robots-david-levy-loebner.
44. Joanna Bryson, « Robots Should Be Slaves », *op. cit*.

Perspective éthique en intelligence artificielle : décoder les biais discriminatoires dans les décisions algorithmiques

Sandrine Charbonneau, *Université Laval*

RÉSUMÉ : Les développements d'algorithmes d'intelligence artificielle (IA) sont porteurs de nombreux bénéfices et probablement d'autant de risques pour la collectivité. Nous croyons qu'ils méritent d'être étudiés et encadrés par une perspective éthique, afin de ne pas reproduire ni renforcer les inégalités et les injustices sociales et économiques déjà présentes. Notre intention sera d'éclairer les risques d'effets discriminatoires dans les décisions algorithmiques en expliquant leur fonctionnement et les sources possibles de leurs biais. Nous verrons qu'un usage de ces technologies fait sans préoccupation pour les membres historiquement plus vulnérables et marginalisés de la société peut rapidement mener à des cas très graves de discrimination. Nous soulignerons qu'il est difficile, mais pas impossible de minimiser ces biais discriminatoires dans les décisions des algorithmes. Au final, nous défendrons que même si l'IA peut nous sembler complexe et opaque dans son fonctionnement, il n'en tient qu'à nous de placer les limites de son utilisation en fonction des valeurs que nous voulons promouvoir.

Technology is neither good nor bad; nor is it neutral. - Melvin Kranzberg

Introduction

Dans la dernière décennie, l'intelligence artificielle (IA) a connu des avancées extrêmement rapides, alors que des algorithmes toujours plus sophistiqués ont été conçus afin d'optimiser nos prises de décisions dans de multiples domaines[1]. On peut notamment observer des applications communes de l'IA dans notre quotidien par la présence de publicités en ligne, dont les contenus sont personnalisés en fonction de nos préférences. Dans bien des cas cependant, les choix effectués grâce à ces programmes sont moins connus, mais peuvent avoir des impacts majeurs dans nos existences. Pensons par exemple à des autorisations de prêts bancaires, des offres d'embauches, ou encore à la durée des peines d'emprisonnement[2]. S'il est complexe de formuler une définition précise et consensuelle de ce qu'est l'IA, cette technologie implique néanmoins les capacités suivantes : « correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation[3] ».

Divers biais[4] peuvent toutefois se glisser à tout moment dans le processus de création algorithmique, entre autres chez les programmeurs et programmeuses, ou encore dans les bases de données qui sont utilisées pour le codage. L'ampleur des impacts que pourrait avoir cette technologie sur nos sociétés et les incertitudes qui persistent quant à notre connaissance de celle-ci devraient nous inciter à adopter une analyse éthique de ses enjeux, afin de minimiser les risques et les dommages que l'IA peut causer aux individus. À cet égard, la programmeuse informatique et activiste Joy Buolamwini décrit notre empressement face aux développements de ces algorithmes en disant : « We have entered the age of automation overconfident, yet underprepared. If we fail to make ethical and inclusive artificial intelligence we risk losing gains made in civil rights and gender equity under the guise of machine neutrality[5] ».

Suivant les propos de Buolamwini, nous discuterons dans cet article des effets potentiellement discriminatoires de la logique prédictive des algorithmes décisionnels d'IA d'apprentissage artificiel. Notre objectif sera de démontrer comment divers biais peuvent se glisser dans les algorithmes d'IA lors de leur programmation

et d'expliquer les effets potentiellement discriminatoires qui peuvent résulter de leur utilisation. Nous soutiendrons que ces technologies peuvent apporter des bénéfices à la société, à condition que celle-ci établisse des balises éthiques qui garantissent que l'utilisation de ces outils « intelligents » ne contribue pas à renforcer les inégalités et les injustices déjà présentes, ce qui n'est pas si aisé à mettre en place. Pour l'expliquer, nous décrirons en premier lieu brièvement le fonctionnement de ce type d'algorithme et relèverons certains de ses apports pour la société, en mettant aussi en garde contre plusieurs dommages qu'il pourrait causer. En second lieu, nous présenterons les types de dommages qui peuvent être causés par des algorithmes biaisés et d'où ces biais discriminatoires peuvent provenir. En troisième lieu, nous mentionnerons deux obstacles majeurs que représente la lutte aux biais, d'abord en ce qui concerne l'aspect technique et ensuite, l'aspect politico-économique. En dernier lieu, nous proposerons une piste de solution afin de réduire la possibilité de biais dans les décisions algorithmiques, passant par des procédures d'encadrement éthique et d'audit.

1. Qu'est-ce que l'IA ?

1.1. Algorithmes d'apprentissage artificiel : prédire et corréler

À l'origine, les approches classiques « symboliques » de l'IA fonctionnaient avec des algorithmes postulant des règles logiques lors de situations précises à partir d'un jeu fini de données d'entraînement[6]. Pensons par exemple à un programme capable de battre à chaque fois les humains aux échecs par sa capacité supérieure de calculer les meilleurs coups possibles. Les progrès les plus récents en IA tiennent surtout du développement de techniques d'apprentissage artificiel avancé sans supervision, tel que l'apprentissage profond (*deep learning*). Ces types d'algorithmes peuvent alors effectuer des corrélations dans des données qui leur permettent d'effectuer des jugements probabilistes plus poussés. Résumons de manière très simplifiée le fonctionnement du modèle de *deep learning* : à partir d'une grande quantité de données, souvent qualifiées de données massives ou de big data (qui peuvent être notamment des chiffres ou des images), ce type d'algorithme d'IA est programmé pour pouvoir

s'entraîner à repérer lui-même des motifs (*patterns*) à travers ces mêmes données et peut, par la suite, lorsque de nouvelles données sont entrées, prédire des résultats. Par exemple, à partir d'une grande base de données d'entraînement comportant des images de chats et de chiens, on peut apprendre à l'algorithme à départager lui-même les différentes espèces. Lorsqu'on lui présenterait une nouvelle image d'un de ces animaux, il pourrait indiquer s'il s'agit de l'une ou de l'autre des espèces (ou aucune des deux), et ce, avec un certain pourcentage d'erreur[7]. Notons qu'il existe de multiples autres modes d'apprentissage artificiel, mais que nous discuterons dans notre article de ceux qui ont une logique probabiliste et corrélationniste, dont les implications en matière d'enjeux éthiques sont similaires.

Ce mode de fonctionnement largement prédictif et corrélationniste est donc ce qui rend l'IA aussi intéressante, mais tout aussi risquée. Comme l'expliquent les sociologues Dominique Cardon et Bilel Benbouzid, « les machines prédictives prétendent calculer les phénomènes sociaux sans s'appuyer sur les attributs catégoriels qui servent ordinairement à enregistrer les acteurs et leurs actions[8] » : puisqu'on laisse l'algorithme « apprendre » par lui-même, il peut effectuer des corrélations que nous n'aurions jamais pu effectuer nous-mêmes, faute de puissance de calcul, mais qui peuvent aussi se révéler erronées à notre sens.

Présentons d'abord les bénéfices que cette technologie représente, pour ensuite exprimer certains des risques qu'elle comporte.

1.2. Les bénéfices

Bien qu'une grande partie des bénéfices entourant les applications de l'IA reste à être prouvée - pour ne nommer qu'un exemple : l'automatisation complète des véhicules qui pourrait entraîner une diminution considérable des accidents sur la route[9]-, plusieurs avantages se font largement sentir. Les résultats les plus spectaculaires de l'utilisation de l'IA se font surtout remarquer dans le domaine de la santé, où certains algorithmes permettent de détecter des maladies avec plus de précision que les meilleurs médecins dans le domaine[10]. L'amélioration des soins de santé due à l'IA est constatée par cette plus grande rapidité et précision des diagnostics,

tout comme par la découverte de nouveaux traitements et par la possibilité d'effectuer un meilleur suivi chez les patients et patientes par l'utilisation de dossiers électroniques[11].

On peut également parler de la croissance économique des entreprises créant et utilisant des technologies d'IA. Depuis les dernières années, celles-ci permettent de générer des milliards de dollars en décuplant la productivité des compagnies, tout en promettant des hausses importantes de profits d'ici les dix prochaines années[12]. Toutes ces entreprises qui utilisent l'IA peuvent optimiser leur efficacité, que ce soit au sein de leur organisation interne, ou encore dans leur offre de produits et de services. L'IA permet aussi une meilleure analyse des opérations financières, aide à la détecter les fraudes[13], facilite la gestion des transports et du secteur agricole, tout comme la prédiction de la météo et de catastrophes naturelles[14].

L'IA promet donc de révolutionner une majorité de secteurs de la société en augmentant l'efficacité et la précision de multiples tâches.

1.3. Risques

Plusieurs individus et organisations ont toutefois commencé à s'intéresser aux principes et aux valeurs qui devraient accompagner notre utilisation et notre développement de l'IA, telle que la *Déclaration de Montréal pour un développement responsable de l'intelligence artificielle*, publiée en 2018[15]. Parmi les principes qui ont été énoncés dans ce document, on retrouve notamment un principe d'équité des êtres humains par rapport aux décisions des algorithmes. Pour démontrer l'importance de telles considérations, prenons le cas d'un algorithme ayant entraîné des effets inattendus. Dans les dernières années, une équipe avait conçu un algorithme programmé pour différencier les chiens de race Husky des loups : les gens qui l'ont codé ont réalisé que l'algorithme se basait plutôt sur le paysage pour distinguer les images de chiens et de loups, ces derniers étant le plus souvent dans la neige. En enlevant les paysages, l'algorithme ne faisait plus la différence entre les deux[16]. Si ce cas peut sembler amusant, d'autres peuvent bien vite devenir profondément discriminatoires.

Joy Buolamwini, qui a la peau noire, a testé plusieurs logiciels de reconnaissance faciale commerciaux fonctionnant avec des algorithmes d'IA entraînés par apprentissage artificiel. Elle a remarqué que ceux-ci ne détectaient pas son visage. Ces logiciels détectaient cependant les visages de ses collègues au teint plus pâle et ont même détecté la présence de son visage lorsqu'elle a mis sur celui-ci un masque blanc[17]. Ces programmes, dans les meilleurs cas, parvenaient généralement à identifier les visages d'hommes blancs avec des taux d'erreurs de quelques pourcentages et d'une dizaine de pourcentages d'erreurs pour les visages de femmes blanches. Dans les pires cas, ils arrivaient à des taux d'erreurs de près de 50 % pour les visages de femmes Noires. En d'autres termes, plus le teint du visage étudié était foncé et plus les traits de ce visage tendaient vers des caractéristiques féminines, plus le taux d'erreurs était élevé. Notons que depuis la publication de cette étude de Buolamwini et de ses collègues, ces compagnies ont révisé leurs programmes et fait diminuer grandement leurs taux d'erreurs (sans pour autant atteindre un degré de précision aussi élevé pour chaque genre et teint de peau)[18].

Le problème avec ces programmes tient surtout du fait qu'ils sont en vente libre, prêts à être utilisés de multiples façons, comme à des fins de sécurité et de vérification de l'identité[19]. Les entreprises ou organismes qui les achètent peuvent alors commettre et reproduire des actions discriminatoires. Si les systèmes d'ouverture de portes d'un bâtiment étaient dotés de l'un de ces programmes, bien des gens pourraient être incapables d'y entrer. Nous verrons dans la prochaine section à quel point ce genre de technologie, avec ses résultats biaisés, peut avoir des effets bien plus préjudiciables sur les individus que ce genre d'agacement.

2. Biais discriminatoires : effets concrets et sources multiples

Nous présenterons dans cette section deux types de dommages, soit d'allocation de ressources et d'opportunités, puis de représentation. Nous poursuivrons avec deux sources de biais liés aux IA, venant d'une part des gens qui programment et de l'autre, des bases de données qui les composent. Le souci de coder

des algorithmes précis et efficaces peut réduire les caractéristiques sociales des humains à de simples données, au détriment du droit de chaque personne à un traitement juste et équitable.

2.1. Types et étendue des dommages

Les dommages pouvant découler des résultats biaisés des algorithmes peuvent être divisés en deux catégories selon Kate Crawford, chercheuse en informatique sur les impacts sociaux des IA : des dommages d'allocation et de représentation[20]. Les problèmes d'allocation sont plus directs, ils influencent par exemple l'octroi de ressources (comme les prêts bancaires et hypothèques), d'opportunités (comme les offres d'embauche et places dans une université) et de services (comme la livraison dans certains secteurs). Ces dommages transactionnels sont plus faciles à quantifier et sont le résultat d'une décision à un temps précis. Les dommages de représentation touchent quant à eux les attitudes et les croyances, reflétant les diverses représentations que des individus peuvent avoir de la société au niveau culturel. Ces derniers sont le plus souvent ignorés en IA, étant plus difficiles à formaliser. Dans cette catégorie, on a par exemple détecté des stéréotypes selon le genre dans les outils de traduction et des outils de reconnaissance faciale ayant des difficultés à identifier les visages de gens non caucasiens[21] : pensons à Google qui avait développé une application de photo ayant étiqueté deux personnes Noires comme étant des singes[22].

De plus, comme le fait remarquer Crawford, les modèles d'algorithmes d'IA fonctionnant par apprentissage artificiel peuvent commettre des erreurs se répandant très vite et à grande échelle. Une même erreur de biais pourrait par exemple toucher jusqu'à un ou deux milliards d'utilisateurs et utilisatrices par jours[23] (pensons simplement à la quantité de gens utilisant des services comme Facebook et Google qui peuvent être affectés si les algorithmes de ces derniers comportent des biais).

2.2. Sources des biais

2.2.1. Les humains et la programmation :

Nombreuses sont les occasions où des algorithmes d'IA ont pris des décisions injustes en discriminant des gens, notamment en fonction de leur genre, de leur race, ou de diverses conditions socio-économiques. Les biais menant à ces discriminations peuvent venir consciemment ou inconsciemment des gens qui programment par l'entretien de préjugés sur divers groupes sociaux, ou par un manque de connaissances et de souci envers des réalités marginalisées. Ces biais s'introduisent alors dans les algorithmes lors des choix de programmation, notamment par les paramètres qui sont sélectionnés pour les configurer.

Cathy O'Neil, une docteure en mathématiques qui travaille dans le domaine des algorithmes, a par exemple déjà interrogé un docteur en statistique qui codait des algorithmes calculant les risques de récidives pour des prisons d'État aux États-Unis. Elle lui a demandé s'il utilisait la « race » comme critère pour coder le programme, ce qu'il a nié. Il a cependant affirmé utiliser les codes postaux, puisqu'ils offrent beaucoup plus de « précision » dans les résultats. Le problème est toutefois que les codes postaux peuvent être de bons indicateurs par « proxy[24] » de la « race » et de la situation économique. L'algorithme, en calculant les risques de récidives, pouvait effectuer des corrélations entre les codes postaux de gens vivant dans des milieux plus défavorisés - dans ce cas, une majorité de gens Noirs - leur assignant des indices de récidives systématiquement plus élevés. O'Neil remarque que plusieurs scientifiques travaillant avec des données massives se voient plutôt comme des techniciens, qui doivent suivre leurs livres et leurs définitions d'optimisation, sans penser aux plus grandes conséquences de leur travail sur la vie et les droits des gens. Selon la mathématicienne, ce raisonnement est le paradigme de la situation actuelle, où certaines valeurs d'efficacité sont plus importantes chez bien des programmeurs et programmeuses que les concepts d'égalité et d'équité[25].

S'il est difficile de se faire un portrait exact de la communauté travaillant à concevoir des IA quant à leur niveau de sensibilisation

par rapport aux impacts sociaux qu'ont leurs algorithmes, plusieurs chercheurs et chercheuses comme O'Neil constatent qu'il reste d'importants efforts de conscientisation sociale à effectuer.

2.2.2. Les humains et les données

Le manque de sensibilisation des humains influençant leurs choix de programmation n'est pas le seul élément en cause dans la présence de biais dans les IA. Celles-ci sont conçues pour apprendre et fonctionner avec les données qui sont à leur disposition : « Ce sont les données, la variété et la qualité de celles-ci qui rendent l'algorithme capable d'un meilleur discernement. Des données peu nombreuses ou relatant des pratiques discriminatoires peuvent reproduire des biais ou en créer, par exemple, en faisant des corrélations entre des éléments qui ne devraient pas être liés[26] ». Un exemple flagrant de discrimination en lien avec l'exemple précédent d'O'Neil a été rapporté par des journalistes de ProPublica, soit celui de l'outil COMPAS, utilisé aux États-Unis pour prédire le risque de récidive des gens accusés de crimes. Dans ce dossier, les journalistes donnent l'exemple de deux crimes similaires, l'un commis par une personne blanche et l'autre par une personne Noire. L'algorithme prédisant leur risque de commettre à nouveau un crime dans le futur donne pourtant un risque plus élevé à la personne Noire, alors qu'elle donne un risque plus faible à la personne blanche. Pour des raisons de choix de critères de calcul dans la programmation, mais aussi et surtout de quantité de données, l'algorithme a « appris » de l'historique des crimes et des sentences passés, reproduisant alors dans ses prédictions les biais racistes des arrestations et des jugements criminels aux États-Unis envers la population Noire[27]. Les scores rapportés par ces outils prédictifs ne sont d'ailleurs pas censés permettre aux juges de donner des sentences plus sévères, mais seulement d'orienter leur jugement. Dans les faits, plusieurs ont cité les résultats de l'algorithme pour justifier leurs décisions[28].

Crawford critique l'approche qu'elle nomme *data fundamentalism*, où corrélation est associée à causalité et où les données massives sont toujours perçues comme offrant des vérités objectives. Selon ses travaux, les données sont souvent prises comme un reflet précis du

monde social, alors qu'il y a toujours des distorsions dans la collecte de données. Celles-ci ne sont jamais « neutres » ou « impartiales », mais peuvent exclure et diviser. Crawford explique que les sciences travaillant avec des Big Data doivent se demander : d'où proviennent leurs données, quelles ont été les méthodes employées pour les analyser et quels sont les biais possibles lors de leur interprétation[29] ? La vigilance est de mise par rapport à la création et l'utilisation de bases de données non représentatives et discriminatoires. Certaines peuvent receler des historiques de pratiques injustes, comme un projet d'algorithme de recrutement d'Amazon, qui désavantageait les candidatures des femmes à l'embauche. Heureusement, il a été décidé que cet algorithme ne serait pas utilisé[30]. Il est important de critiquer les données qui seront utilisées pour entraîner l'IA et de comprendre d'où elles viennent, afin de prévoir quelles discriminations elles pourraient engendrer.

3. Obstacles pour minimiser les biais

Nous montrerons en quoi les algorithmes d'IA posent de sérieux enjeux éthiques d'explicabilité lors de leur utilisation, par un obstacle technique et un obstacle politico-économique. D'une part, la complexité des IA fonctionnant par apprentissage artificiel (et plus particulièrement par apprentissage profond) rend leurs résultats très opaques. De l'autre, les algorithmes développés par les entreprises sont actuellement protégés par les droits commerciaux de propriétés intellectuelles et de propriété privée, ce qui rend souvent impossible la possibilité d'examiner et de critiquer leurs décisions.

3.1. Boîte noire : manque de transparence et d'explicabilité des décisions

Le problème de la boîte noire (*black box*) est celui où on peut observer les entrées (*inputs*) et les sorties (*outputs*) dans un programme informatique, mais sans bien comprendre comment les uns ont pu mener aux autres[31]. Comme l'expliquent les auteurs du rapport Villani de 2018, des algorithmes d'IA fonctionnant par apprentissage profond déploient tant de calculs et de paramètres qu'il est « presque impossible de suivre le cheminement de l'algorithme

de classement[32] », ce qui peut rendre leurs résultats peu ou pas explicables. Ci-dessous, voici une schématisation de ce à quoi peut ressembler un « arbre décisionnel » informatique très simple, représentant quelques couches de calculs des « réseaux neuronaux artificiels » de l'algorithme[33].

Avec des milliers, voire des millions de variables et d'entrées, il peut être pratiquement impossible de retrouver les origines des biais discriminatoires dans ces technologies. Même en testant l'algorithme, on ne pourrait pas savoir, par exemple, si les discriminations viennent plutôt de la programmation, des données d'entraînement, ou même des deux.

Le rapport Villani donne l'exemple d'algorithmes de Google de ciblage publicitaire, qui ont eu tendance à proposer aux femmes des offres d'emplois à plus faible rémunération que celles qu'ils proposent aux hommes, ou encore d'algorithmes qui ont tenté de prédire les secteurs les plus susceptibles de criminalité aux États-Unis, ayant pour résultat l'augmentation de la surveillance dans des quartiers pauvres à prédominance afro-américaine[34]. Ces cas ayant été constatés, on pourrait dans le premier établir que ces biais viennent, par exemple, uniquement de l'historique des données utilisées, avec des statistiques montrant des salaires moins élevés pour les femmes. Toutefois, on pourrait aussi penser que ces traitements discriminatoires proviennent des préjugés entretenus par les gens ayant programmé les algorithmes, qui auraient pu inconsciemment proposer certains types d'emplois moins bien rémunérés aux femmes et non aux hommes. Le problème reste donc la difficulté d'établir

d'où proviennent les biais, en raison de l'opacité des algorithmes. En examinant les codes qui les composent, on ne saurait départager avec précision quelles données et quels choix ont pu donner tels résultats.

Bien que la plupart des pays où ces algorithmes sont utilisés possèdent des lois pour protéger les populations des discriminations, on ne peut établir avec certitude le poids qu'ont les prédictions des IA dans la prise de décisions aux effets discriminatoires. Comme dans le cas de l'outil COMPAS mentionné plus haut, les juges ne devraient pas se baser uniquement sur les algorithmes pour déterminer les sentences, mais ils et elles le font parfois. À plus long terme, si l'utilisation d'algorithmes d'aide à la décision continue à se répandre, sans être plus transparents dans leurs résultats, nous pourrions décider de ne plus leur confier certaines tâches pour éviter le maintien et la reproduction d'injustices[35].

3.2. Entreprises privées : maximisation des profits et de l'efficacité

Un obstacle supplémentaire se dresse à la possibilité de comprendre les décisions des algorithmes d'IA et de critiquer leurs biais le cas échéant : le fonctionnement de ces algorithmes est considéré comme un secret d'affaires, les formules qui les composent sont donc protégées par les droits de propriété intellectuelle et privée, empêchant qui que ce soit d'analyser en détail les étapes de leurs calculs[36]. Dans bien des cas, le souci de générer plus de profits passe outre le respect des humains pouvant être affectés par les indications des algorithmes. Plusieurs compagnies cachent délibérément les résultats de leurs modèles, avec des hordes d'avocats et de lobbys pour les défendre, comme Google, Amazon et Facebook, dont les algorithmes seuls valent des centaines de millions de dollars[37].

Les domaines de la finance et des assurances sont particulièrement reconnus pour avoir massivement recours à des algorithmes afin d'évaluer leur clientèle actuelle ou future. Une énorme quantité de données est récoltée sur ces gens, avec leur consentement ou à leur insu[38]. Comme dans le cas des programmes d'IA prédisant les risques de récidives chez les criminels et criminelles, les algorithmes

de finance et d'assurance affectent les scores de fiabilité des individus en fonction d'informations directement pertinentes (par exemple, le respect du Code de la route) et d'autres par « proxy », dont les liens de causalité sont très discutables (comme le code postal résidentiel du conducteur ou de la conductrice, ou encore les régions visitées en voiture, suivies par géolocalisation)[39]. De nombreuses personnes ont critiqué ces scores opaques, qualifiés par exemple d'« arbitrary, and discriminatory[40] », mais sans obtenir de garanties de ces compagnies ou des gouvernements que la situation soit étudiée.

Cette réalité concrète illustre bien que malgré la conscientisation montante dans le domaine de l'IA sur les impacts sociaux des algorithmes d'apprentissage artificiel, il semble manquer de mesures concrètes de contrôle et d'évaluation de ces logiciels quant aux impacts de leur utilisation.

4. Pistes de développement et d'encadrement

Faute de pouvoir développer longuement sur les multiples façons dont pourraient être encadrés les algorithmes d'IA, nous ne ferons qu'esquisser une piste de solution aux problèmes précédemment énoncés. Nous avancerons que les biais indésirables se retrouvant dans des programmes informatiques peuvent être limités par une réglementation très serrée des IA, impliquant des procédures d'audit, où sont testés les algorithmes avant et après leur mise en marche. L'objectif serait de s'assurer qu'à chaque étape, nous puissions avoir une compréhension suffisante de leurs résultats et que nous nous assurions que ceux-ci ne soient pas discriminatoires. Notre conclusion sera que si les algorithmes démontrent des biais dans leurs décisions, ils devraient être révisés ou sinon, interdits.

4.1. Audits et réglementation

Face aux défis techniques et juridiques que posent l'opacité et le secret d'affaires des algorithmes d'IA, nous proposons d'adopter des réglementations exigeant plus de transparence et un suivi serré des compagnies programmant et utilisant ces algorithmes. Nous demandons pour ce faire que ces derniers puissent être sujets à des examens et à des audits avant et après leur mise en application. Il

va sans dire que nous contestons les lois empêchant l'accès aux codes des algorithmes pour des raisons de propriété privée ou intellectuelle. Nous proposons la mise en place de comités d'analyse et de révisions indépendants, soumis à des clauses de confidentialité. Par exemple, un groupe pourrait être composé d'experts et d'expertes en informatique et de spécialistes en éthique qui n'ont pas de conflits d'intérêts avec les compagnies développant et utilisant l'algorithme étudié.

Nos suggestions, à l'instar des recommandations du rapport AI Now 2017, se concentrent principalement sur les tâches des IA reliées aux organismes publics dont les décisions peuvent avoir des impacts majeurs sur la vie des gens, par exemple dans les domaines de la justice criminelle, de la santé, de l'aide sociale et de l'éducation[41]. Comme nous l'avons mentionné, la complexité des algorithmes d'IA d'apprentissage automatique rend très difficile, voire impossible la compréhension claire de leurs résultats. Nous ne nous attendons donc pas à une explicabilité complète du programme dans les résultats fournis, mais nous tenons à la possibilité d'interroger ses bases de données, pour mieux comprendre le fonctionnement de l'algorithme.

De plus, avant d'utiliser des données pour entraîner un modèle, il faut s'assurer de pouvoir comprendre d'où viennent ces données et ce qu'elles représentent, en s'interrogeant notamment sur les méthodes employées pour les collecter. Comme le soutient Crawford, il faut se tourner vers les sciences sociales pour cette partie du travail et aller plus loin que de demander « combien » aux gens, mais aussi « pourquoi » et « comment »[42]. Il est important que cette étape d'analyse des données soit effectuée avant d'entraîner des programmes pouvant rendre des décisions lourdes de conséquences. Une fois que les algorithmes auront été testés exhaustivement et que leur utilisation ne semblera pas mener à des conclusions injustement discriminatoires[43], il faudrait s'assurer que leurs résultats soient les plus transparents que possible pour ceux et celles qui les interprètent.

Revenons à l'exemple d'Amazon, où l'entreprise a décidé d'elle-même de ne pas utiliser son algorithme pour les embauches, puisqu'il défavorisait les candidatures des femmes[44]. Dans une autre situation,

il a été révélé par Bloomberg que cette même compagnie n'offrait pas son service de livraison en 24 heures dans certains quartiers à majorité afro-américaine aux États-Unis, les algorithmes utilisant les codes postaux pour optimiser les possibilités de profits. À la suite de la publication de l'article sur cette réalité discriminatoire, les maires de New York, Boston et Chicago ont critiqué Amazon, qui a rapidement décidé d'offrir le service à ces endroits précédemment exclus. Cependant, aucune réglementation ne les obligeait à le faire et aucune loi ne les empêchait concrètement de concevoir et d'utiliser ces algorithmes discriminatoires[45].

Selon nous, dans des cas plus délicats (comme pour juger des sentences de prison), les machines ne devraient pas être autorisées à rendre une décision automatique ou une suggestion sans autorisation ou examen humain. Nous privilégierons toujours des approches plus complètes avec un jugement humain, même si celles-ci s'avèrent plus coûteuses en temps, puisqu'un gain en efficacité pourrait se solder par de graves préjudices. Si l'algorithme, après avoir été analysé et testé sous les conseils venant de disciplines aux points de vue variés, ne nous permet pas d'affirmer que ses résultats sont exempts de biais causant de la discrimination, il ne devrait pas avoir le droit d'être utilisé.

5. Conclusion

Dans cet article, nous avons tenté d'expliquer le fonctionnement de certains types d'IA, ainsi que les bénéfices potentiels de son utilisation, tout comme certains risques qu'elle comporte pour la société. Nous avons voulu démontrer de quelle manière des biais aux effets discriminatoires peuvent être introduits dans les algorithmes lors de la programmation, par les préjugés ou le manque de sensibilité de ceux et celles qui programment envers certains groupes de personnes, ou encore dans leur choix des données utilisées. Nous avons décrit quelques-uns des nombreux obstacles techniques et politiques se dressant contre les tentatives d'amenuiser ou d'annuler les effets discriminatoires des biais de l'IA. Finalement, nous avons lancé une piste de réflexion sur les possibilités d'encadrer

la conception et les usages des algorithmes d'IA, par des procédures d'audit et de réglementation.

Nous sommes encore aux débuts du développement de ces nouvelles technologies, au moment où les lois n'ont pas encore balisé ni encadré l'ensemble de leur fonctionnement. Comme l'avancent la chercheuse en « gouvernementalité algorithmique » Antoinette Rouvroy et d'autres , les algorithmes d'IA commencent à s'étendre dans toutes les sphères de notre quotidien et sont appelés à jouer des rôles de plus en plus grands au sein de la gouvernance même de la société[46]. Les impacts de ces outils de décision basés sur des prédictions devraient être réfléchis sous des points de vue inclusifs et non discriminatoires. Il nous semble le respect les droits de chaque être humain à être traité justement et équitablement devraient guider de façon prioritaire nos choix futurs. Nous pensons que malgré le fait que les jugements humains puissent être défaillants, il faut trancher sur les cas où ils restent préférables aux décisions des machines, puisqu'il est plus facile de comprendre et de critiquer des raisonnements humains dans des cas de litiges.

Nous croyons tout de même que l'essor du développement en IA peut se faire de manière responsable. Il comporte de nombreuses opportunités d'améliorer notre qualité de vie, malgré les risques de reproduction et de maintien de discriminations. Sur le plan des bénéfices possibles, nous pouvons aussi mentionner la possibilité d'adopter des algorithmes moins biaisés que les humains sur certains plans. Des juges peuvent avoir tendance à accorder des peines d'emprisonnement inéquitables en raison de biais racistes, mais aussi pour des raisons aussi banales que l'heure à laquelle ils et elles rendent leur sentence - une étude ayant démontré que les peines étaient plus sévères avant la pause dîner en raison de leur faim[47]. Si la création d'algorithmes purement « neutres » ou sans biais est impossible, on peut toutefois tenter d'en programmer qui soient moins partiaux et plus justes dans leurs décisions.

1. Nous nous concentrerons sur les enjeux touchant l'intelligence artificielle « simple » ou « faible », soit les programmes informatiques d'algorithmes

de calcul actuels et non l'intelligence artificielle « complexe » qui pourrait être autant, sinon plus intelligente que l'humain, comportant selon plusieurs des risques existentiels. Pour plus de détails sur cette dernière, voir par exemple : Nick Bostrom et Eliezer Yudkowsky, « The ethics of artificial intelligence », dans Keith Frankish *et al.* [dir.], *The Cambridge Handbook of Artificial Intelligence*, Cambridge, Cambridge University Press, 2014, p. 316-334.

2. Notons que l'IA n'est pas systématiquement utilisée dans les situations mentionnées en exemple. Il peut être difficile, voire impossible pour le public de savoir si les décisions rendues par des systèmes informatiques résultent de l'utilisation de l'IA ou non. C'est toutefois une pratique en croissance dans la plupart des entreprises : Louis Columbus, « 10 Charts That Will Change Your Perspective On Artificial Intelligence's Growth », dans *Forbes,* [En ligne], https://www.forbes.com/sites/louiscolumbus/2018/01/12/10-charts-that-will-change-your-perspective-on-artificial-intelligences-growth/ (Page consultée le 6 juillet 2019).
3. Andreas Kaplan et Michael Haenlein, « Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence », dans *Business Horizons*, vol. 62, no° 1, janvier 2019, p. 15.
4. Nous emploierons le terme « biais » dans le sens large de « biais cognitif », incluant toute erreur de raisonnement (que ce soit dans la théorie ou pratique, de façon consciente ou inconsciente). Nous mettrons l'accent sur les biais qui poussent vers la partialité et les préjugés : Marie van Loon, « Biais cognitifs (version académique) » dans *M. Kristanek* [dir.], L'Encyclopédie Philosophique, 2018.
5. Maureen McElaney, « Cognitive Bias in Machine Learning », The Data Lab [En ligne], https://medium.com/codait/cognitive-bias-in-machine-learning-d287838eeb4b (Page consultée le 9 juillet 2019).
6. Jocelyn Maclure et Marie-Noëlle Saint-Pierre, « Le nouvel âge de l'intelligence artificielle : une synthèse des enjeux éthiques », dans *Les cahiers de propriété intellectuelle*, vol. 30, no° 3, octobre 2018, p. 748.
7. SAS Institute, « Deep Learning: What it is and why it matters » [En ligne], https://www.sas.com/en_us/insights/analytics/machine-learning.html (Page consultée le 6 juillet 2019).
8. Bilel Benbouzid et Dominique Cardon, « Machines à prédire », dans *Réseaux*, no° 5, 2018, p. 28.

9. Waymo, « Waymo Safety Report. On The Road To Fully Self-Driving » [En ligne], https://waymo.com/safety/ (Page consultée le 11 juillet 2019).
10. Martin Stumpe et Craig Mermel, « Improved Grading of Prostate Cancer Using Deep Learning », Google AI Blog [En ligne], https://ai.googleblog.com/2018/11/improved-grading-of-prostate-cancer.html (Page consultée le 6 juillet 2019).
11. Lisa Morgan, « Artificial Intelligence in Healthcare: How AI Shapes Medecine », *Datamation* [En ligne], https://www.datamation.com/artificial-intelligence/artificial-intelligence-in-healthcare.html (Page consultée le 11 juillet 2019).
12. Louis Columbus, *op.cit*.
13. Santana Wilson, « 3 Ways AI is Used in Business Process Optimization », Oracle Data Science [En ligne], https://www.datascience.com/blog/ai-for-business-process-optimization, (Page consultée le 13 juillet 2019).
14. Magnimind Academy, « Invaluables Societal Benefits of AI », Medium [En ligne], https://becominghuman.ai/invaluable-societal-benefits-of-ai-2ed62f7a653f (Page consultée le 12 juillet 2019).
15. Université de Montréal, « Déclaration de Montréal pour un développement responsable de l'intelligence artificielle » [En ligne], https://www.declarationmontreal-iaresponsable.com/ (Page consultée le 6 juillet 2019).
16. Marco Tulio Ribeiro *et al*., « "Why Should I Trust You?" : Explaining the Predictions of Any Classifier », dans *arXiv:1602.04938* [cs, stat], février 2016, p. 9.
17. Joy Buolamwini et Timnit Gebru, « Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification », dans *Journal of Machine Learning Research,* vol. 81, février 2018.
18. Deborah Raji Inioluwa et Joy Buolamwini, « Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products », dans *Conference on Artificial* Intelligence, Ethics, and Society, 2019.
19. Face++, « Face Detection » [En ligne], https://www.faceplusplus.com/face-detection/ (Page consultée le 13 juillet 2019).
20. Il existe bien sûr d'autres manières de les traiter, celle de Crawford nous apparaissant la plus complète au niveau des biais ayant des effets discriminatoires dans les décisions prises par des algorithmes.
21. Kate Crawford, « The Trouble with Bias », Conférence prononcée au NIPS 2017, Longbeach, CA, le 5 décembre 2017, 17:00 [En ligne], https://www.youtube.com/watch?v=fMym_BKWQzk, (Page consultée le 30 novembre 2018).
22. Jana Kasperkevic, « Google says sorry for racist auto-tag in photo app », *The Guardian* [En ligne], https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app, (Page consultée le 12 juillet 2019).
23. Kate Crawford, *op. cit.*, 5:30.
24. La discrimination par « proxy » est une forme de discrimination par facteurs indirects : « We formalize a notion of proxy discrimination in data-driven systems, a class of properties indicative of bias, as the presence of protected class correlates that have causal influence on the system's output » : Anupam Datta *et al.*, « Proxy Discrimination in Data-Driven Systems Theory and Experiments with Machine Learnt Programs », dans *arXiv:1707.08120v1*, juillet 2017, p. 1.
25. Tom Upchurch, « To work for society, data scientists need a hippocratic oath with teeth », *Wired UK* [En ligne], https://www.wired.co.uk/article/data-ai-ethics-hippocratic-oath-cathy-o-neil-weapons-of-math-destruction (Page consultée le 18 décembre 2018).
26. Jocelyn Maclure et Marie-Noëlle Saint-Pierre, *op. cit.*, p. 756.
27. Aliya Saperstein, *et al.* « The Criminal Justice System and the Racialization of Perceptions », dans T*he ANNALS of the American Academy of Political and Social Science*, vol. 651, no° 1, janvier 2014, p. 104-121.
28. Julia Angwin *et al.*, « Machine Bias », ProPublica [En ligne], https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Page consultée le 23 novembre 2018).
29. Kate Crawford, « The Hidden Biases in Big Data », Harvard Business Review [En ligne], https://hbr.org/2013/04/the-hidden-biases-in-big-data (Page consultée le 23 novembre 2018).
30. Cathy O'Neil, « Amazon's Gender-Biased Algorithm Is Not Alone », Bloomberg [En ligne], https://www.bloomberg.com/opinion/articles/2018-10-16/amazon-s-gender-biased-algorithm-is-not-alone (Page consultée le 18 décembre 2018).
31. Frank Pasquale, *The Black Box Society*: *The Secret Algorithms That Control Money and Information*, Cambridge, Harvard University Press, 2015, p. 3.
32. Cédric Villani, *et al*., « Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne », France, mars 2018, p. 142 [En ligne], https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf (Page consultée le 18 décembre 2018).

33. Jonathan Shaw, « Artificial Intelligence and Ethics », Harvard Magazine [En ligne], https://harvardmagazine.com/2019/01/artificial-intelligence-limitations (Page consultée le 12 juillet 2019).
34. Cédric Villani, *op.cit.*
35. *Ibid.*
36. Kate Crawford, « Artificial Intelligence's White Guy Problem », *The New York Times* [En ligne], https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html (Page consultée le 24 septembre 2019).
37. Cathy O'Neil, Weapons of math destruction : *how big data increases inequality and threatens democracy*, New York, Crown, 2016, p. 29.
38. Autre enjeu éthique majeur à propos duquel nous ne pourrons développer.
39. Cathy O'Neil, *op. cit.* p. 164-171.
40. Frank Pasquale, *op. cit.*, p. 23.
41. Alex Campolo, *et al.*, The AI Now Report : T*he Social and Economic Implications of Artificial Intelligence,* AI Now Institute, 2017, p. 1.
42. Kate Crawford, « The Hidden Biases in Big Data », *op. cit*.
43. Nous n'avons pas l'espace nécessaire pour expliquer de quelles manières nous pensons que les tests devraient être conduits. Nous croyons, à l'instar de Cathy O'Neil, qu'une suggestion intéressante pour conduire des audits sur les boîtes noires de programmes d'IA serait de les tester avec des cas variés, afin de s'assurer qu'ils ne mènent pas systématiquement à des résultats discriminatoires. (Tom Upchurch, *op. cit.*).
44. Cathy O'Neil, « Amazon's Gender-Biased Algorithm Is Not Alone », *op. cit*.
45. David Ingold et Spencer Soper, « Amazon Doesn't Consider the Race of Its Customers. Should It? », *Bloomberg* [En ligne] http://www.bloomberg.com/graphics/2016-amazon-same-day/, (Page consultée le 23 novembre 2018).
46. Marc-Olivier Bherer, « En 2018, résistez aux algorithmes avec la philosophe Antoinette Rouvroy », *Le Monde* [En ligne], https://www.lemonde.fr/idees/article/2017/12/29/en-2018-resistez-aux-algorithmes-avec-la-philosophe-antoinette-rouvroy_5235555_3232.html (Page consultée le 20 décembre 2018).
47. Shai Danziger *et al.*, « Extraneous factors in judicial decisions », *Proceedings of the National Academy of Sciences,* vol. 108, no^{o} 17, avril 2011, p. 6889-6892.

Polarisation des opinions et délibération démocratique : l'influence des algorithmes

Éric Gagnon, *Université Laval*

RÉSUMÉ : Depuis la montée des réseaux sociaux dans la vie démocratique, on pourrait s'attendre à une démocratisation des tribunes publiques ainsi qu'à un développement d'une délibération saine. Toutefois, la réalité est toute autre. En effet, les réseaux sociaux sont devenus un lieu de polarisation des opinions. Une meilleure compréhension de la psyché humaine et de l'usage contre-productif de l'Intelligence artificielle (IA) qui régule les réseaux sociaux pourra permettre d'expliquer le phénomène de radicalisation des croyances, ce à quoi nous nous consacrerons dans cet article. Nous chercherons également à trouver des solutions aux problèmes qu'amène l'usage actuel des algorithmes de l'IA sur les réseaux sociaux.

Introduction

Depuis les élections présidentielles américaines de 2016, on peut avoir l'impression qu'un fossé idéologique divise profondément la société américaine. Un rapport publié à la suite d'une vaste étude menée par le *Pew Research Center* révèle que ce fossé entre les partisans démocrates et républicains est bien réel, qu'il est apparu il y a une vingtaine d'années, et s'est creusé encore davantage entre 2014 et 2017 (voir Annexes 1 et 2). En effet, le rapport du *Pew Research* Center montre que la médiane idéologique des partisans démocrates et républicains s'est déplacée vers les extrêmes, suivant un phénomène de radicalisation politique. Durant cette période, les partisans démocrates sont devenus plus libéraux tandis que les républicains sont devenus plus conservateurs. De plus, le scandale bien connu de *Cambridge Analytica* impliquant la plateforme *Facebook* a révélé que les algorithmes des réseaux sociaux influencent les pensées et les choix des électeurs. Cette polarisation

des opinions peut être expliquée par l'influence des réseaux sociaux, et plus précisément par les algorithmes utilisés par l'Intelligence artificielle qui régissent ceux-ci. Mon but sera donc de comprendre et d'expliquer cette influence de l'Intelligence artificielle sur la polarisation des opinions, en tenant compte du fonctionnement particulier de la raison humaine.

Pour ce faire, je m'appuierai d'abord sur une nouvelle théorie prometteuse, issue des sciences cognitives du raisonnement, portant sur la nature et le fonctionnement de la raison, pour fournir une compréhension du potentiel épistémique de la raison humaine. Cette compréhension du rôle et du fonctionnement de la raison permettra de cerner et d'expliquer le phénomène de la polarisation des opinions. J'aborderai finalement ce phénomène dans le contexte des réseaux sociaux, pour montrer que la polarisation des opinions est la conséquence de l'usage actuel de l'Intelligence artificielle et que cela nuit au potentiel épistémique de la raison.

1. Le potentiel épistémique de la raison

1.1. Les limites épistémiques de la raison

Dans *The Enigma of Reason*, Hugo Mercier et Dan Sperber, chercheurs français en sciences cognitives du raisonnement, présentent une nouvelle théorie portant sur la raison : la *théorie argumentative* du raisonnement. Ils proposent dans cet ouvrage une vision interactionniste de la raison : c'est-à-dire que celle-ci aurait évolué chez les êtres humains dans un contexte social, et jouerait un rôle important au sein de leurs interactions.

Comme toute autre espèce du règne animal, les êtres humains font des inférences au quotidien[1]. Mercier et Sperber définissent l'*inférence* comme « the extraction of new information from information already available, whatever the process[2] ». Ces informations déjà disponibles proviennent de régularités empiriques, c'est-à-dire de régularités dont nous faisons l'expérience. Par exemple, lorsque je marche dehors et qu'il se met à pleuvoir, je serai trempé. En faisant cette expérience plusieurs fois (régularité), j'en arrive à inférer que chaque fois qu'il se met à pleuvoir, je serai trempé.

Par ailleurs, nous avons rarement conscience du cheminement inférentiel qui nous mène à une conclusion. Cette dernière se présente soudainement à notre esprit : pour reprendre notre exemple, lorsqu'il se met à pleuvoir, j'ai la certitude que je serai mouillé sans devoir pour cela référer à toutes les expériences qui m'ont permis de tirer cette conclusion. Une conclusion consciente qui est fruit de mécanismes inférentiels inconscients est ce que Mercier et Sperber nomment une *intuition*. Chaque intuition se présente à notre esprit avec un sentiment de confiance fort ou faible. Plus nous revivons souvent une expérience, plus l'intuition qui en est issue est renforcée, et plus elle nous semble évidente et vraie. Il en va de même pour nos opinions et nos croyances, qui sont des types d'intuitions.

C'est lorsque l'on doit défendre et partager nos intuitions que la raison a un rôle à jouer. Pour nos auteurs, elle s'est développée au cours de l'évolution de l'espèce humaine afin de faire face à deux obstacles majeurs, et ce avec ses deux fonctions principales : « [one] function helps solve a major problem of coordination by producing justifications. The other function helps solve a major problem of communication by producing arguments[3] ». Dans le cadre de notre recherche, nous nous pencherons sur la seconde fonction, celle de produire des arguments en vue de convaincre nos pairs de se rallier à nos intuitions. Pour cela, la raison se sert de *son processus inférentiel particulier* qu'est le *raisonnement*. Ce dernier a deux capacités : produire des arguments et les évaluer. Pour ce faire, dans les deux cas, il extrait de nouvelles informations de celles déjà disponibles. Puisque ces nouvelles informations sont des contenus conscients qui surviennent soudainement à notre esprit, elles sont des intuitions[4]. Bref, pour appuyer nos croyances, qui sont des intuitions, le raisonnement produit des arguments à l'aide de nouvelles intuitions.

En ce qui concerne la seconde capacité, le raisonnement permet d'évaluer les arguments. Il s'agit de la *vigilance épistémique* de la raison, c'est-à-dire *l'attention portant sur le poids et la validité des arguments* permettant de les accepter ou de les rejeter. En effet, puisque les arguments s'appuient sur des intuitions qui se présentent à notre esprit avec un degré faible ou fort de confiance, leur apparente validité dépend du même type de confiance. Par exemple, pour

quelqu'un qui a vu la rougeole disparaître avec l'arrivée des vaccins, cela constitue un argument évident de leur efficacité. Ce n'est toutefois pas un argument aussi convaincant pour quelqu'un qui n'a pas ce vécu et l'intuition qui en est inférée.

Dans un contexte de dialogue, le raisonnement permet au communicateur de convaincre autrui de ses croyances ou opinions en lui présentant des arguments ; pour le destinataire, le raisonnement a pour rôle d'évaluer les arguments qui lui sont présentés. Le destinataire assure donc une forme de vigilance épistémique[5]. En ce sens, « [reasoning] involves two capacities, that of producing arguments and that of evaluating them. [...] Jointly they constitute [...] one of the two main functions of reason and the main function of reasoning : the argumentative function[6] ».

Mercier et Sperber ajoutent toutefois que dans la réalisation de ces deux fonctions, la production d'arguments et leur évaluation, un biais cognitif joue un rôle important : *le biais du parti pris* (*myside bias*). En effet, ce biais a été mis en lumière dans une étude réalisée par Deanna Kuhn en 1991. Dans cette recherche, on demandait aux participants de donner et de justifier leur opinion sur certains sujets. S'ils argumentaient sans difficulté en faveur de leurs croyances, seulement 14% des participants étaient en mesure de produire des contre-arguments à leur propre opinion. De la même manière, il nous est difficile d'argumenter contre une opinion que nous partageons. Le biais de parti pris nous empêche donc d'être critiques à l'égard de nos opinions, ou de celles que nous partageons. De plus, une autre étude, menée par Victoria F. Shaw en 1996, a démontré que les individus n'ont aucune difficulté à trouver des contre-arguments aux idées et arguments qui s'opposent aux leurs[7]. Ainsi, comme le suggèrent Mercier et Sperber :

> [What] these results—and many others[8]—show is that people have no general preference for confirmation. What they find difficult is not looking for counterevidence or counterarguments in general, but only when what is being challenged is their own opinion. Reasoning does not blindly confirm any belief it bears on. Instead, reasoning systematically works to find reasons for our ideas and against ideas we oppose. It always takes our side. As a result, it is preferable to speak of a myside bias rather than of a confirmation bias[9].

L'étude réalisée par Kuhn a aussi démontré que lorsque la raison produisait des arguments, ces derniers étaient souvent faibles (phénomène qui nous est familier dans les commentaires sur internet)[10]. Par exemple, un participant expliquait l'échec scolaire par le manque d'un certain nutriment et, dans les mots de Kuhn, « [the participant] makes it clear that the existence of the phenomenon itself is sufficient evidence that it is produced by the cause the [participant] invokes[11] ». Autrement dit, un argument très faible suffisait au participant pour justifier son opinion, car le biais du parti pris rend la raison paresseuse (*lazy*). Elle relâche sa vigilance épistémique, c'est-à-dire l'attention portée au poids et à la validité des arguments qu'elle mobilise lorsqu'elle défend sa propre opinion, mais demeure exigeante à l'égard des contre-arguments avancés[12].

À l'image d'un avocat qui prend le parti de son client, pour la raison, tous les arguments sont bons pour défendre son intuition. Mais seuls les meilleurs arguments seront acceptés par le jury, soit la raison d'autrui. C'est dans cette différence entre la faible vigilance épistémique envers nos arguments et la vigilance plus élevée envers les opinions qui divergent des nôtres, que résident les bénéfices de l'argumentation et le rôle social de la raison :

> [The] most difficult task, finding good reasons, is made easier by the myside bias and by sensible laziness. The myside bias makes reasoners focus on just one side of the issue rather than having to figure out on their own how to adopt everyone's perspective. Laziness lets reason stop looking for better reasons when it has found an acceptable one. The interlocutor, if not convinced, will look for a counterargument, helping the speaker produce more pointed reasons. By using bias and laziness to its advantage, the exchange of reasons offers an elegant, cost-effective way to solve a disagreement[13].

Ainsi, dans le contexte du dialogue, la raison peut utiliser le biais du parti pris à son avantage pour accomplir ses deux principales

fonctions : la production d'arguments et leur évaluation. En effet, c'est dans un échange d'arguments que les meilleurs ressortent, et que seuls ceux-ci sont acceptés en fin de compte, puisque la paresse que l'on a à l'égard de nos propres arguments est contrebalancée par l'exigence élevée de l'interlocuteur soutenant une position inverse.

Nous avons mentionné plus haut que la raison a évolué chez l'être humain en raison de l'utilité sociale que ses fonctions remplissent. Dans une perspective évolutionniste, chaque espèce développe des traits qui lui permettent de tirer profit de l'environnement spécifique dans lequel elle vit. Nous devons donc maintenant nous pencher sur l'environnement qui a favorisé l'évolution de la raison, et qui permet la réalisation de ses fonctions.

1.2. L'efficacité de la raison en délibération

La raison humaine est adaptée à un environnement social particulier : un contexte de dialogue et de délibération. En effet, « the normal conditions for the use of reasons are social, and more specifically dialogic. Outside of this environment, there is no guarantee that reasoning acts for the benefits of the reasoner[14] ». C'est donc seulement dans ce genre de contexte délibératif que la raison s'avère être un avantage évolutif. Hugo Mercier et Hélène Landemore définissent par ailleurs la délibération comme étant « an activity [...] to the extent that reasoning is used to gather and evaluate arguments for and against a given proposition[15] ». Autrement dit, la délibération oppose au moins deux croyances qui sont défendues et attaquées à l'aide d'arguments tout en étant soumises, l'une et l'autre, à l'évaluation de leur interlocuteur. C'est donc dans une hétérogénéité d'intuitions se confrontant que le potentiel de la raison se réalise. En effet, c'est ce que soutient une étude réalisée en 1997, sous la direction de Deanna Kuhn, qui a révélé que des élèves produisaient de meilleurs arguments pour défendre leurs opinions à la suite d'un débat dans un contexte de délibération, dans lequel sont échangés des arguments[16].

Non seulement la délibération de groupe donne aux participants la chance de réviser leurs croyances, mais elle permet aussi de faire ressortir de nombreux arguments forts en vue de peser chaque

point de vue, et mène donc à de meilleurs résultats épistémiques. En effet, pour Mercier et Landemore, « [when] people are engaged in a genuine deliberation, the [...] bias present in each individual's reasoning is checked, compensated by the [...] bias of individuals who defend another opinion[17] ».

Dans le cas où les opinions se polarisent, c'est pourtant l'inverse qui se produit : les positions de chaque parti sont plus extrêmes après un débat qu'avant celui-ci. Or, nous observons que ce phénomène est exacerbé sur les plateformes des réseaux sociaux. C'est pourquoi il nous semble important de comprendre le contexte particulier qu'ils offrent à la délibération, ainsi que le rôle particulier qu'y tiennent les algorithmes.

2. La polarisation comme échec du potentiel de la raison

Pour comprendre le phénomène de la polarisation des opinions sur les réseaux sociaux, nous pouvons nous tourner vers les analyses de Cass R. Sunstein. Selon ce chercheur, trois facteurs peuvent amener un individu à radicaliser ses opinions : l'information disponible, la confiance en ses croyances et l'influence sociale[18]. Dans le premier cas, une personne peut être exposée à des informations en faveur de ses croyances préétablies qui les renforcent et les déplacent vers une position plus extrême. Dans le second cas, il s'agit du fait que plus nous gagnons confiance en notre opinion, plus elle se renforce. De plus, comme le souligne Sunstein, la « confiance augmente si d'autres semblent partager le même point de vue[19] ». Plus précisément, plus nous sommes entourés de personnes qui partagent notre croyance, plus nous prenons confiance en elle. Une confiance trop élevée mènerait à une croyance inébranlable. Finalement, l'influence sociale peut aussi mener quelqu'un à se radicaliser, par exemple lorsque des individus adoptent les croyances qui dominent dans le groupe auquel ils appartiennent[20]. Dans ce dernier cas, il faut considérer les liens affectifs (familiaux, amicaux, etc.) : plus ces liens sont forts, moins il y a de désaccords entre les individus du groupe et plus l'influence sociale est forte.

Ces trois cas de figure ont en commun de se produire lorsque les croyances d'un groupe sont homogènes. En effet, lorsque les seuls

arguments que l'on entend à *répétition* sont ceux qui soutiennent la croyance dominante de notre groupe, cette croyance et l'intuition sur laquelle elle repose sont sans cesse renforcées. Ainsi, il semble, à première vue, que la délibération peut, dans certains contextes, mener à une polarisation des opinions et remettre, ce qui remet en cause la démocratie délibérative elle-même[21].

Toutefois, rappelons que l'environnement favorable à la réalisation des fonctions de la raison est un environnement caractérisé par une hétérogénéité d'intuitions. En effet, pour Mercier et Landemore, « [p]olarization and overconfidence happen because not all group discussion is deliberative[22] ». En somme, un groupe aux intuitions homogènes n'est précisément pas délibératif, puisqu'il ne peut mener à des désaccords, à des échanges considérant les pour et les contre d'une proposition : il ne peut donc pas réaliser le potentiel de la raison. Ainsi, Sunstein partage la conclusion de Mercier, de Sperber et de Landemore : la polarisation des opinions est le fruit d'un groupe homogène. En effet, en reprenant les thèses de tous ces auteurs, on peut comprendre que le biais du parti pris affaiblit la vigilance épistémique de la raison à l'égard des arguments fournis pour justifier une intuition partagée par le groupe. En effet, Sunstein ajoute que « de nombreux extrémistes souffrent d'une "paralysie épistémologique" parce qu'ils ne connaissaient qu'une toute petite fraction de ce qu'il y a à savoir[23] », car ils se privent des bénéfices épistémiques de la délibération. Comme le disent Mercier et Sperber :

> [If] people have their ideas closely aligned to start with, it leads to polarization. When people start with conflicting ideas and no common goal, it tends to exacerbate differences. Group discussion is typically beneficial when participants have different ideas and a common goal[24].

Pour créer un environnement délibératif, il faut donc une diversité d'opinions, mais aussi un but commun. Un but commun peut être par exemple la résolution d'une énigme scientifique, la réduction de l'impact des changements climatiques, la diminution des inégalités socioéconomiques, etc. Or, ces questions sont précisément celles qui animent nos démocraties et doivent être résolues par une délibération

démocratique, caractérisée par un échange d'arguments. Dans un contexte où les réseaux sociaux facilitent les échanges entre les individus, il est à se demander si de telles plateformes contribuent à la délibération démocratique ou plutôt à la polarisation des opinions.

3. Les réseaux sociaux et la délibération démocratique

3.1. Les réseaux sociaux comme lieu de polarisation

Comme nous venons de le voir, les milieux homogènes sont des environnements qui ne permettent pas de réaliser le potentiel de la raison parce qu'ils ne favorisent pas la délibération. Or, les réseaux sociaux se révèlent être des environnements de ce genre, homogènes et non délibératifs, entre autres en raison de l'utilisation l'intelligence artificielle (IA) et des algorithmes. Par IA, nous entendons : « a system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaption[25] ». L'IA est ainsi un système qui utilise les données à sa portée afin de mener à bien un but qui lui est fixé. Pour arriver à un tel but, elle utilise des algorithmes, c'est-à-dire, « a set of rules that precisely defines a sequence of operations such that each rule is effective and definite and such that the sequence terminates in a finite time[26] ». Il s'agit plus précisément d'un ensemble de règles d'une précision telle qu'elle peut cibler une séquence avec les données qui sont à sa disposition. En somme, l'IA est un ensemble d'algorithmes, qui ciblent des séquences en vue d'un but qui lui est fixé. Par exemple, dans le contexte de l'utilisation des réseaux sociaux, elle peut cibler des individus en vue de les exposer à certains contenus selon leur affiliation idéologique.

Un réseau social, tel que Facebook, qui utilise une IA en vue de servir ses intérêts, peut donc influencer ce qui peut apparaître dans le fil d'actualité de ses utilisateurs. En ce sens, l'IA est mobilisée pour afficher aux utilisateurs ce qu'ils souhaitent voir, grâce au traitement des données effectuées par les algorithmes. Comme Sunstein le soutient, nous pouvons aussi bien nous servir des réseaux sociaux pour diversifier nos sources d'information que pour ne voir que ce qui s'accorde à nos opinions[27]. Dans ce dernier cas, l'IA peut réguler le flux d'information présenté à l'utilisateur de

telle manière qu'il soit homogène, ce qui aura tendance à renforcer et à radicaliser ses idées. Toutefois, comme ce fut le cas dans le scandale de *Cambridge Analytica*, l'IA peut être appelée à influencer de manière partisane ce même flux d'information pour amener les utilisateurs à se mobiliser pour un parti plutôt qu'un autre[28].

En 2016, 68 % des Américains possédaient un compte Facebook[39] et, la même année, 45 % prenaient leurs nouvelles et suivaient l'actualité sur cette plateforme[30]. On peut donc prédire que les partisans démocrates ou républicains, s'ils suivent leurs membres élus au Congrès sur Facebook et qu'ils prennent leurs nouvelles d'actualité chez ces derniers, deviendront, sous l'effet de leurs publications sur le réseau social, polarisés dans leurs opinions. Or, il est probable que c'est qui est en train de se passer actuellement, puisque des statistiques témoignent d'un fort déplacement de la médiane démocrate vers une idéologie plus libérale, entre 2014 et 2017 (voir Annexes 1 et 2). Ce déplacement coïncide avec la forte augmentation des publications de députés démocrates sur leur page Facebook depuis l'élection de Donald Trump à la présidence américaine en 2016 (voir Annexe 3). De plus, une analyse des publications sur Facebook des membres du Congrès américain, révèle que :

> Democrats expressed political opposition nearly five times as much under Trump as they did during the last two years of Barack Obama's presidency [and] Members of Congress who expressed political opposition most often were also the most liberal or conservative[31].

Non seulement les députés démocrates exprimaient davantage que par le passé leur opposition politique sur Facebook, mais les députés ayant les positions les plus extrêmes (du côté démocrate comme du côté républicain) étaient aussi ceux qui publiaient le plus souvent et qui possédaient le plus grand nombre d'abonnés[32]. En ce sens, la polarisation des opinions des partisans démocrates entre 2014 et 2017 pourrait en partie s'expliquer par l'influence des réseaux sociaux. Si l'on considère que des partisans démocrates font partie de près de la moitié des Américains qui prennent leurs

nouvelles sur Facebook et qu'ils suivent peut-être entre autres la page de leurs députés, ces derniers auraient pu contribuer à polariser leurs électeurs en publiant davantage de contenu marqué par l'opposition politique aux républicains. Quant au Parti républicain, sa faible quantité de publications exprimant une opposition politique depuis 2015, en baisse depuis l'élection de Trump, semble concorder avec l'absence de déplacement de la médiane de ses partisans depuis 2014 (voir Annexes 1, 2 et 3).

3.2. Une délibération à l'ère numérique est-elle possible ?

L'influence polarisatrice des réseaux sociaux et de leurs algorithmes nous apparaît donc claire. La question qui se pose maintenant est de savoir si une délibération démocratique demeure possible sur les réseaux sociaux. J'avancerais que oui, mais à condition qu'il y ait un changement dans l'usage des algorithmes. En effet, les espoirs de Mercier et de Landemore visent plutôt sur les institutions que les citoyens[33]. Considérant la nature de la raison humaine et les types d'environnement l'influençant tant négativement que positivement, il semble que le rôle des institutions est important puisqu'elles constituent un type d'environnement. De plus les réseaux sociaux sont des institutions à la fois sociales et politiques, puisqu'elles sont des tribunes permettant aux citoyens de s'exprimer et de délibérer entre eux.

Vu le potentiel épistémique limité des humains et de leur raison en contexte solitaire ou homogène, un changement d'usage des algorithmes de Facebook, qui ferait des réseaux sociaux un environnement plus hétérogène, permettrait de réduire, sans éradiquer, la polarisation des opinions. Les algorithmes devraient donc être utilisés de façon à nous exposer à des idées divergentes des nôtres et présentées avec une argumentation soutenue. Toutefois, l'information et les faits devront venir de sources crédibles, qui méritent notre confiance quant à la véracité de ce qui est rapporté ou avancé par elle. Par exemple, ces sources crédibles peuvent être des organismes scientifiques qui ne sont pas rattachés à des entreprises qui auraient des gains à faire à pervertir les résultats. De plus, ces sources devraient défendre leur contenu de manière rigoureuse.

De cette manière, un échange délibératif adéquat pourrait avoir lieu. L'approche peut sembler idéaliste, mais considérant que les IA peuvent trouver des contenus homogènes, rien ne les empêche de trouver des contenus dont les points de vue diffèrent du point de vue de l'utilisateur. Si des sources fiables sont citées et vérifiées par l'IA, elles pourraient être favorisées dans l'apparition du fil d'actualité de ces derniers. Néanmoins, jusqu'à ce que de tels changements soient appliqués, il faudrait exhorter les utilisateurs à s'exposer volontairement aux points de vue inverses tant que ces derniers soient soutenus de manière solide et rigoureuse.

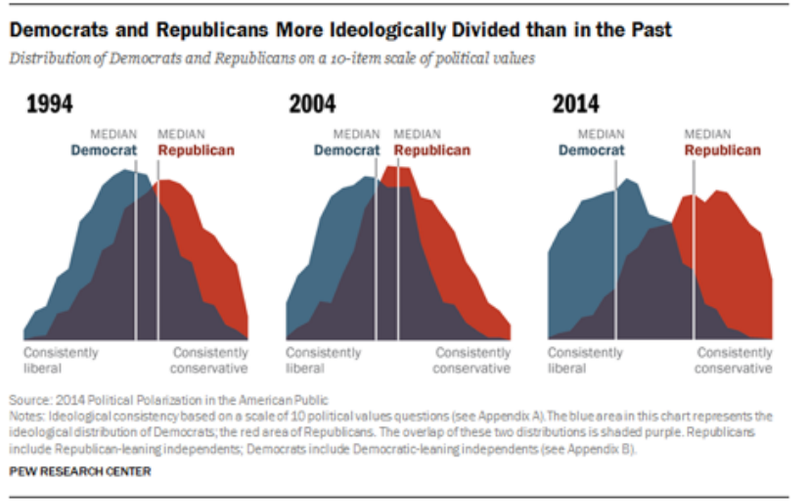
Considérant que les entreprises possédant des réseaux sociaux ne semblent pas portées par elles-mêmes à de tels changements, on peut placer nos espoirs concernant une telle réforme institutionnelle dans une intervention de l'État qui pourrait prendre la forme d'un projet de loi imposant aux entreprises d'exposer leurs utilisateurs à des opinions divergentes. En effet, considérant le bien-être de nos démocraties, il semble important d'éviter d'exacerber les différences et les tensions. À une époque où les sociétés sont de plus en plus pluralistes et où le contact entre gens de croyances divergentes est plus fréquent, les bénéfices épistémiques de la délibération seraient encore plus profitables qu'à tout autre moment de l'histoire de l'humanité.

Conclusion

Nous avons discuté du potentiel épistémique des êtres humains avec la théorie argumentative du raisonnement de Mercier et de Sperber qui nous a permis de comprendre comment les croyances peuvent survenir à l'esprit humain et comment la raison les défend. De plus, nous avons vu comment une personne peut utiliser la raison pour évaluer les arguments d'autrui et changer d'idée si de meilleurs arguments lui sont présentés. En nous basant sur cette théorie, nous avons étudié comment la polarisation témoigne de l'échec du potentiel de la raison humaine et comment elle survient. Nous avons ensuite montré que les réseaux sociaux sont des vecteurs de polarisation des opinions. Nous avons ensuite semé des pistes de réflexion concernant les moyens de changer l'usage des algorithmes

et des IA utilisées par les réseaux sociaux en vue d'aider à la délibération démocratique, et non lui nuire. Toutefois, il n'en demeure pas moins que les individus utilisant les réseaux sociaux doivent en faire un usage responsable, c'est-à-dire s'exposer à des points de vue divergents de leurs opinions. En outre, nous préconisons un encadrement législatif de l'usage des algorithmes responsables du contenu présenté sur les réseaux sociaux, afin de maintenir un environnement démocratique hétérogène et sain.

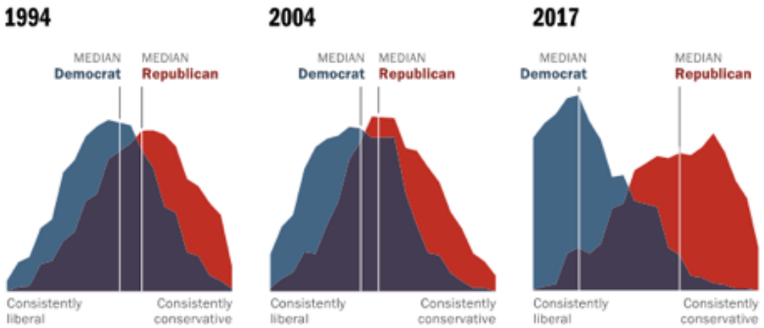
Annexe 1[34]



Annexe 2[35]

Democrats and Republicans more ideologically divided than in the past

Distribution of Democrats and Republicans on a 10-item scale of political values



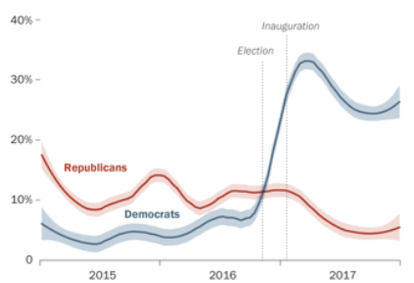
Notes: Ideological consistency based on a scale of 10 political values questions (see methodology). The blue area in this chart represents the ideological distribution of Democrats and Democratic-leaning independents; the red area of Republicans and Republican-leaning independents. The overlap of these two distributions is shaded purple.
Source: Survey conducted June 8-18, 2017.

PEW RESEARCH CENTER

Annexe 3[36]

Following Trump's election, Facebook posts from Democrats in Congress included more oppositional language

Average % of posts expressing political opposition



Note: Political opposition includes statements that oppose President Trump or Republicans and conservatives (for Democrats) and statements that express opposition to President Obama, Hillary Clinton or Democrats and liberals (for Republicans). Lines are based on LOWESS estimates. The shaded regions are the 95% confidence bands for the estimated trends.
Source: Pew Research analysis of Facebook posts created by members of Congress between Jan. 1, 2015 and Dec. 31, 2017.
"Taking Sides on Facebook: How Congressional Outreach Changed Under President Trump"

PEW RESEARCH CENTER

1. Hugo Mercier et Dan Sperber, *The Enigma of Reason*, Cambridge, Harvard University Press, 2017, p. 53-57.
2. *Ibid*., p. 53.
3. *Ibid.*, p. 183.
4. *Ibid*., p. 131-132.
5. *Ibid.*, p. 193-195.
6. *Ibid*., p. 198-199.
7. *Ibid*., p. 218.
8. Voir à ce propos Hugo Mercier et Dan Sperber, « Why do humans reason ? Arguments for an argumentative theory » dans *Behavioral and Brain Sciences,* vol. 34, no° 2, 2011, p. 57-74.
9. Hugo Mercier et Dan Sperber, *The Enigma of Reason*, *op. cit.*, p. 218.
10. *Ibid*., p. 223.
11. Deanna Kuhn, *The Skills of Argument*, Cambridge, Cambridge University Press, 1991, p. 87.
12. Mercier, Hugo et Dan Sperber, *op. cit*., p. 235.
13. *Ibid*., p. 236.
14. *Ibid*., p. 247.
15. Hugo Mercier et Hélène Landemore, « Reasoning is for Arguing: Understanding the Successes and Failures of Deliberation, *Political Psychology* », vol. 33, no° 2, 2012, p. 251.
16. Deanna Kuhn, Victoria Shaw et Marc Felton. « Effects of dyadic interaction on argumentative reasoning », dans *Cognition and Instruction*, vol. 15, no° 3, p. 287-315 [en ligne], https://www.researchgate.net/profile/Mark_Felton2/publication/239959892_Effects_of_Dyadic_Interaction_on_Argumentive_Reasoning/links/571e43ef08aeaced7889deb6/Effects-of-Dyadic-Interaction-on-Argumentive-Reasoning.pdf. Cité dans Mercier, Hugo et Dan Sperber, op. cit., p. 228.
17. Hugo Mercier et Hélène Landemore, *op. cit.*, p. 22.
18. Cass R. Sunstein, « Délibération, nouvelles technologies et extrémisme », trad. Solange Chavel, dans *Raison publique,* 2016 [en ligne], http://www.raison-publique.fr/article776.html.
19. *Ibid.*
20. *Ibid.*
21. *Ibid.*
22. Hugo Mercier et Hélène Landemore, *op. cit.*, p. 20.
23. Cass R. Sunstein, *op. cit.*

24. Hugo Mercier et Dan Sperber, *op. cit.*, p. 334.
25. Andreas Kaplan et Michael Haenlein, « Siri, Siri, in my hand: Who's the fairest in the land ? On the interpretations, illustrations, and implications of artificial intelligence », dans *Business Horizons*, vol. 62, no° 1, 2019, p. 15-25 [en ligne], https://www.sciencedirect.com/science/article/pii/S0007681318301393 ?via%3Dihub.
26. Herbert L. Stone, *Introduction to Computer Organization and Data Structures*, New York, Éditions McGraw-Hill, 1972, p. 8.
27. Cass R. Sunstein, *op. cit*.
28. Lisa Maria Neudert et Nahema Marchal, « Polarisation and the use of technology in political campaigns and communication », dans *European Parliamentary Research Service,* 2019, p. 23 [en ligne], http://www.europarl.europa.eu/RegData/etudes/STUD/2019/634414/EPRS_STU(2019)634414_EN.pdf.
29. Aaron Smith et Monica Anderson, « Social Media Use in 2018: A majority of Americans use Facebook and YouTube, but young adults are especially heavy users of Snapchat and Instagram » dans *Pew Research Center,* mars 2018, p. 2.
30. Elisa Shearer et Jeffrey Gottfried, « News Use Across Social Media Platforms 2017 », dans *Pew Research Center*, septembre 2017, p. 6.
31. Patrick Van Kessel, Adam Hughes et Solomon Messing. « Taking Sides on Facebook: How Congressional Outreach Changed Under President Trump: Democratic legislators' opposition on Facebook spiked after Trump's election, while angry reactions increased among all congressional Facebook followers », dans *Pew Research Cente*r, juillet 2018, p. 3-4.
32. Adam Hughes et Onyi Lam, « Highly ideological members of Congress have more Facebook followers than moderates do », dans *Pew Research Center,* août 2017 [en ligne], https://www.pewresearch.org/fact-tank/2017/08/21/highly-ideological-members-of-congress-have-more-facebook-followers-than-moderates-do/.
33. Hugo Mercier et Hélène Landemore, *op. cit.*, p. 25.
34. Michael Dimock, Carroll Doherty, Jocelyn Kiley et Russ Oates, « Political Polarization in the American Public: How increasing Ideological Uniformity and Partisan Antipathy Affect Politics, Compromise and Every Life », dans *Pew Research Center*, juin 2014, p. 6.
35. Carroll Doherty, Jocelyn Kiley et Bridget Johnson, « The Partisan Divide on Political Values Grows Even Wider: Sharp shifts among Democrats on aid to needy, race, immigration », dans *Pew Research Cente*r, octobre 2017, p. 12.
36. Patrick Van Kessel, Adam Hughes et Solomon Messing, *op. cit.*, p. 3.

Politique éditoriale de la revue *Phares*

Tous les textes reçus font l'objet d'une évaluation anonyme par les membres du comité de rédaction selon les critères suivants : clarté de la langue, qualité de l'argumentation ou de la réflexion philosophique et accessibilité du propos. En outre, il est important pour les membres du comité que les textes soumis pour un dossier prennent en charge la question proposée de manière effective. La longueur maximale des textes recherchés est de 7000 mots, **notes comprises**. Les textes soumis à la revue doivent respecter certaines consignes de mise en page spécifiées sur le site Internet de la revue, au *www.revuephares.com*. Un texte qui ne respecte pas suffisamment les consignes de mise en page de la revue ou qui comporte trop de fautes d'orthographe sera renvoyé à son auteur avant d'être évalué, pour que celui-ci effectue les modifications requises. Les auteurs seront informés par courriel du résultat de l'évaluation.

Les propositions doivent être acheminées par courriel à l'adresse électronique *revue.phares@fp.ulaval.ca* avant la prochaine date de tombée. Toute question concernant la politique éditoriale de la revue *Phares* peut être acheminée au comité de rédaction à cette même adresse. Les auteurs sont incités à consulter le site Internet de la revue pour être au fait des précisions supplémentaires et des mises à jour éventuellement apportées à la politique éditoriale en cours d'année.