THIERRY MÉNISSIER

CONFIANCE EN L'INTELLIGENCE ARTIFICIELLE ET AUTORITÉ DES MACHINES

1. *De l'usage des technologies d'intelligence artificielle à l'autorité des machines*

Nous voulons porter l'attention sur l'intérêt qu'il y a, en partant des cas offerts par l'observation de l'usage des outils technologiques contemporains, à étudier les relations de confiance envers les « agents non-humains » que sont ces outils et grâce à cela à conceptualiser l'autorité qu'ils sont en train d'acquérir[1]. Nous avons conscience du caractère apparemment iconoclaste du rapprochement que, *via* la thématique de la confiance, nous opérons entre le système technique global aujourd'hui nommé « intelligence artificielle » (IA) et la dimension morale et politique de la notion d'autorité. S'il semble contre-intuitif, ce rapprochement n'en est pas moins nécessité par l'observation de la réalité. Et il est de toute évidence de type problématique, mais présente à ce titre une importance majeure aussi bien pour une éthique des technologies que pour une philosophie politique contemporaine.

Dans cette contribution, l'intelligence artificielle (IA) est entendue *lato sensu*, et renvoie à des réalités variées, telles que le travail « ordinaire » des algorithmes nourris par les méga-données (*big data*) fournies *via* des usages directs (dans les domaines de la mobilité, du rapport à l'énergie, de l'activité sportive, etc.), ou agrégées à distance (et souvent à l'insu) des usagers *via* leur activité numérique, ou encore les diverses interactions possibles avec la robotique (notamment avec les robots non-humanoïdes, aujourd'hui très nombreux). Cet élargisse-

[1] Ce texte est le fruit du travail scientifique qui est mené dans le cadre de la chaire « éthique & IA » soutenue par l'institut pluridisciplinaire en intelligence artificielle MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

ment de la notion d'IA est volontaire : s'il s'agit de partir de cas d'usage, toujours localisés dans des pratiques sociales précises et appelant la description et la caractérisation attentives, le propos doit assumer une certaine de forme de généralité, car celle-ci paraît nécessaire pour le travail scientifique. En effet, ce dernier se situe à l'intersection entre le déploiement réel des technologies et la compréhension systémique de leur sens. Or cette intersection est assurée, au cœur même des usages, par la confiance manifestée par les usagers à l'égard des technologies, confiance qu'il convient à la fois de penser génériquement et de modéliser selon des catégories adéquates. Par-là, l'apport des nouvelles technologies pour la société et pour l'humain doit être questionné en des termes sociaux aussi bien que technologiques, et c'est de cette mise en relation que peut naître une « éthique de l'IA » qui n'est pas une simple régulation des usages *ad hoc* mais l'effet d'une réflexion de fond qui opère à partir de catégories éthiques fondamentales (telles que responsabilité, autorité, liberté)[2].

Fortement interdisciplinaire, une telle démarche sera complète si elle articule trois niveaux différents et complémentaires, à savoir, la recherche en philosophie morale et politique, les études et développements techniques menés par des informaticiens et roboticiens, et enfin celles de chercheurs en psychologie cognitive et sociale, voire, de par la présence des réseaux numériques, celles des chercheurs en sciences de l'information et de la communication. Ainsi qu'on le montrera dans les pages qui suivent, en effet, ce dernier niveau concerne en effet, si l'on dit les choses de manière schématique, l'appréhension expérimentale de la confiance envers les dispositifs de l'IA et leur expertise fondée sur les *data* ; le deuxième, la transformation effective des outils technologiques en agents socio-technologiques ; la première, l'analyse de la possible appréhension du pouvoir des algorithmes comme celui d'autorités dignes de confiance, et l'évaluation d'une telle disposition en termes d'enjeux pour la liberté humaine.

[2] Voir par exemple Floridi (2013), Floridi & Cowls (2019).

2. *Quelles modalités de la confiance à travers les usages de l'IA ?*

Les outils contemporains engendrent certains affects par leur seul usage, il est aisé de le constater. Premièrement, on peut dresser le constat que le déploiement contemporain de l'IA requiert certaines formes de confiance envers les machines, partagées par leurs concepteurs et leurs usagers. Deuxièmement, celui que ces formes de confiance apparaissent indispensables en tant que « ciment » des usages garantissant l'efficacité des machines. On pourrait fournir à l'appui de ce double constat un nombre important d'exemples : aujourd'hui, l'interaction avec un « dialogueur » ou « agent conversationnel » (*chatbot*) sur le site d'une organisation publique ou privée, ou une opération numérique quelconque (réservation d'un séjour ou d'un moyen de transport, achat en ligne de produits et de services, virement bancaire, etc.) réalisée via une application dédiée, demain les déplacements grâce à des véhicules dit « autonomes », ou l'emploi massif de robots domestiques pour la santé ou le soin des personnes, ou la sécurité grâce à des dispositifs biométriques généralisés, ou encore le dialogue personnel avec un assistant-coach numérique, tous ces usages avérés ou émergents impliquent que l'on ait confiance dans la capacité de machines à effectuer ce qu'on leur demande. En 2015, la « Une » de *The Economist* titrant que la *blockchain*, ce genre de système technique complexe déjà en place dans la finance depuis de nombreuses années, est synonyme de « *Trust Machine* » avait été fortement discutée dans les médias (*The Economist* 2015). Le point important de discussion vient du fait que, paradoxalement, la confiance se trouve à la fois sollicitée et mise en question dans ce type de technologie (De Filippi & Wright 2018). A cet égard, c'est l'ensemble des usages de l'IA qui se trouve caractérisé par cette mise en question : l'exemple de la *blockchain* pourrait quasiment valoir comme un cas d'espèce pour les autres technologies de l'IA, machines reposant sur le principe de l'apprentissage qui les rend automatiques de manière exponentielle. A mesure qu'elles apprennent et deviennent « autonomes », ces machines semblent en effet requérir et obtenir une confiance de plus en plus forte de la part des usagers.

Cette situation n'est pas entièrement nouvelle, ainsi qu'on peut l'établir grâce à l'histoire et à la philosophie des techniques : d'autres périodes historiques ont en effet connu des changements techniques radicaux tel que celui que l'intelligence artificielle semble en train d'opérer actuellement, en particulier lors des changements de paradigmes énergétiques (par exemple, lors du passage de la motricité par la vapeur à celle par l'électricité et par les énergies fossiles[3], ou même, en adoptant encore davantage de recul, au moment du « choix du feu » comme paradigme global[4]). De tels changements impliquent toujours qu'un lien de confiance s'établisse envers les nouvelles technologies, soit qu'on dise qu'elles doivent faire preuve de leur fiabilité avant d'être adoptées, soit que l'on considère que les intérêts puissants qui les promeuvent réussissent à faire croire à leur efficience et à leur fiabilité[5].

Aujourd'hui, avec l'essor des systèmes d'IA, la situation apparaît toutefois plus originale car cette confiance s'adresse moins à des outils inertes qu'il est aisé de considérer comme neutres qu'à des dispositifs « intelligents », c'est-à-dire en effet plus ou moins autonomes et susceptibles d'une interaction socialisée. Le déploiement contemporain des intelligences artificielles est en effet compris dans des usages. Or, cette notion, « les usages » concerne l'interaction entre les instruments techniques et les pratiques humaines auxquelles ils donnent lieu, ils combinent l'ergonomie des instruments et leurs emplois variés par des humains qui, pour se les approprier, sont stimulés par le marketing (notamment le « marketing émotionnel » qui, par le biais d'interfaces habiles, joue sur l'impression de socialisation spontanée des instruments[6]).

C'est dans un tel contexte que la question se pose de savoir s'il s'agit dans ces usages d'une réinvention authentique de la confiance ou d'une nouvelle forme de relation[7]. La question se pose d'autant plus que la nature de la confiance s'avère d'elle-

[3] Ainsi que le montre Mitchell (2013).
[4] Voir Gras (2007).
[5] Pour l'analyse de telles dynamiques voir notamment Pestre (2014).
[6] Voir Illouz (2006) ; Martin-Juchat (2014).
[7] Voir Taddeo (2010 a & b : 2011)

même peu claire[8]. Elle se pose également parce, selon l'approche traditionnelle, la notion n'en est pas spontanément employée pour traiter de la relation aux êtres inanimés. Dans une contribution importante pour la compréhension anthropologique de cet affect, le sociologue Louis Quéré formulait mêmes certaines réserves :

> Il n'est pas sûr qu'il y ait un sens à parler de confiance lorsqu'il s'agit de se fier à la stabilité de l'environnement ou à la régularité de comportement de ses objets. En effet, le cas paradigmatique de la confiance est celui d'une relation de confiance entre deux personnes. Les traits caractéristiques de la confiance se maintiennent-ils hors de ce contexte ? Il est possible que nous commettions un abus de langage lorsque nous parlons de faire confiance à un objet ou à une institution (Quéré 2001 : 131).

L'auteur relayait ici un des présupposés souvent et depuis longtemps affirmé par les recherches en sciences sociales[9]. Et en effet, peut-on réellement parler de confiance lorsqu'on se fie à un ensemble de machines automatiques, éventuellement coordonnées entre elles (comme par exemple dans le cas des vastes ensembles que sont les *smart cities*) et dont le fonctionnement global est opaque pour la grande majorité des usagers ?

Dans le même temps, il faut relever une certaine confusion sémantique, lorsque, dans notre actualité, des experts réunis par la puissance publique contribuent à alimenter l'idée d'une mise en relation nécessaire entre le système technique de l'IA et la notion de confiance. A cet égard, la thématique médiatique très puissante autour de l'IA « digne de confiance » (*Trustworthy Artificial Intelligence*) amplifie la confusion en faisant apparaître la confiance dont l'IA peut être investie comme un des nœuds de l'interrogation éthique d'aujourd'hui. En avril 2019, un groupe de 52 experts réunis par la Commission européenne avait publié ses « règles éthiques pour une IA de confiance »[10]. Leur rapport final indique plusieurs critères estimés

[8] Voir Origgi (2008) et Hunyadi (2020).
[9] Voir ces contributions fondamentales de Gambetta (1988) et de Seligman (1997).
[10] Voir European Commission (2019).

nécessaires pour estimer cette confiance, tels que la supervision humaine, la transparence, la robustesse et la non-discrimination, etc. On remarque le caractère hétérogène de la liste fournie, hétérogénéité qui repose sur le mélange entre des critères purement techniques validant un usage fiable de la technologie basée sur la robustesse de cette dernière, et des critères plutôt éthiques basés sur les valeurs de la démocratie (telle que la non-discrimination). Il est de ce fait permis d'estimer que ce premier travail remis par les experts réunis par la Commission Européenne recouvre de si considérables enjeux de réassurance et de sécurisation psychologiques qu'à regarder les choses de près, la notion de confiance ne s'en trouve pas fondamentalement éclairée.

Plusieurs questions se posent en effet de manière plus précise à partir du moment où l'on restitue un certain nombre de nuances sémantiques importantes : d'abord, la fiabilité avérée d'un outil technique aussi automatisé qu'une IA représente certes à la fois la condition de possibilité de sa diffusion dans la société et le résultat de son usage efficace ; mais est-elle assimilable à de la confiance au sens complet du terme ? Ensuite, l'assurance impliquée par la fiabilité de cette machine à travers ses usages, derrière la supposée réinvention de la confiance, à partir de quel seuil produit-elle un « lâcher-prise » de l'autonomie des usagers, où l'humain serait tenté d'abandonner à l'intelligence artificielle une partie de ses prérogatives, avec d'autant plus de facilité que celle-ci paraît le soulager de ses rôles embarrassants (par exemple, dans le cas de la *blockchain*, administrateur et politique, banquier et notaire – autant de rôles à forte contrainte en termes de responsabilité) ?

3. *La conception de l'IA à travers des outils techniques qui deviennent des agents non-humains de la relation sociale*

Pour commencer à répondre à ces questions, il est nécessaire d'approfondir l'hypothèse selon laquelle la « captation de confiance » qui favorise l'usage des technologies de l'IA concerne moins des outils considérés par leurs concepteurs comme des objets neutres que des entités volontairement personnalisées voire personnifiées. De nos jours, les interfaces senso-

rielles des applications et des outils robotiques procèdent en effet toujours en fonction des usages, selon la logique que nous avons évoquée plus haut, c'est-à-dire en stimulant, au sein même de l'emploi des objets, des affects et des symboles. Pour le dire rapidement, leur ergonomie est moins fonctionnelle que sensorielle-culturelle, ou plutôt elle est fonctionnelle parce qu'elle est sensorielle-culturelle. Une telle manière de procéder est mise en œuvre dans le design même des objets, c'est-à-dire dès leur conception, puisque celle-ci s'effectue via l'approche dite « centrée clients » (*Customer centric*) ou plus généralement en référence à celle dite « expérience utilisateurs » (*User experience* ou *UX*)[11]. Et en tant que telle, elle est susceptible de produire des effets en termes d'acceptation ou acceptabilité des technologies, en tout cas elle y contribue efficacement.

Ce constat permet d'abord de valider au contact des chercheurs informaticiens et roboticiens la thèse (aujourd'hui déjà classique pour les sciences humaines et sociales) selon laquelle il n'y a, pour l'outil technique, aucune revendication possible de neutralité axiologique : non seulement la technique comme la science sont liées à des « intérêts » (Voir Habermas 1990), mais encore, en fonction des usages variés auxquels elle donne lieu, et sous l'effet des forces sociales qui en se l'appropriant lui donnent vie, la technologie constitue un puissant moyen de transformation de la réalité (Voir Feenberg 1999). Il convient même d'admettre que « l'objet » technique n'est jamais purement « objectal » : ainsi que l'établissent l'anthropologie et la philosophie des techniques, parce qu'il est à la fois socialisé et socialisateur, il représente toujours un instrument fondamental de médiation, tant entre l'humain et son milieu (Leroi-Gourhan 1964, 1973) qu'entre les individus composant le groupe humain (Simondon 2012). Ces considérations conduisent ensuite à prendre acte comme d'une hypothèse du fait que le monde de l'IA – cette expression désignant le système technique global qui est en train de devenir mondialement dominant -, lorsqu'on l'appréhende du point de vue des usages, n'est nullement désincarné. Il est au contraire peuplé d'entités nouvelles à qui les usagers adressent leurs attentes. Bien qu'encore mal qualifiées

[11] Joët (2000) ; Akrich & Méadel (2004).

sur le plan de l'anthropologie ou de l'ontologie, ces entités ne sont déjà plus des « outils » mais représentent déjà des « amis », des « auxiliaires » ou des « assistants » : la science sociale commence à les identifier comme des agents non-humains de la relation humaine[12].

On a pu récemment écrire qu'il convenait de « sortir du débat ontologique » afin de s'inscrire dans « sociologie pragmatique des interactions entre humains et êtres artificiels intelligents » (Borelle 2018). Si une telle proposition nous paraît intéressante, ce n'est pas que la voie de l'ontologie n'est pas intéressante, c'est que la compréhension ontologique de ces nouveaux êtres passe en effet par l'interaction avec eux, en tant qu'ils sont des agents sociaux (affectivement et symboliquement adressés par leurs concepteurs à leurs usagers potentiels). S'il y a de l'«intelligence » (au sens humain du terme) qui se joue effectivement grâce aux usagers des algorithmes, n'est-ce pas parce qu'il existe de l'interaction entre agents humains et agents non-humains[13] ? Si bien qu'un des caractères fondamentaux de la confiance apparaît dans les échanges, à savoir, l'espoir de socialisation qu'on met dans le lien entre les êtres (Marzano 2010).

4. *Pouvoir de la technologie, influence des agents non-humains : quelle autorité pour l'IA ?*

Via les services qu'ils rendent et l'influence qu'ils exercent déjà sur les comportements, les agents non-humains exercent une action réelle qu'il convient de documenter et de conceptualiser. Si particulière qu'elle soit, il existe de la confiance envers l'action efficace des machines. Nous voulons à présent examiner l'hypothèse que cette confiance conditionne l'essor d'un nouveau genre d'autorité lié à l'efficience technologique.

La dynamique contemporaine d'évolution technique en informatique est notamment polarisée par l'essor de la décision automatisée, par le biais de l'apprentissage automatique (*machine learning*). Des études ont montré que, alors qu'ils escomptent des effets favorables de cette tendance, les usagers mani-

[12] Voir par exemple Caronia & Mortari (2015).
[13] Dumouchel & Damiano (2016).

festent paradoxalement à son propos une faible conscience ; et aussi que, second paradoxe, lorsqu'elle est consciente à leur esprit, elle apparaît assez peu acceptée[14]. Cela signifie qu'il y a bien un « enjeu de pouvoir » dans le rapport tissé entre efficience technique et idéologie promotrice de la technique (Sfez 2002) et confirme que l'action des algorithmes, en étant profondément sociale (Beer 2017), redessine les rapports entre société et pouvoir (Durante 2019). Des conceptualités nouvelles et importantes sont même déjà nées quant à la nature du nouveau pouvoir algorithmique[15]. L'angle de vue que nous adoptons pour notre part vise à confirmer en quoi ce pouvoir peut se voir légitimé au-delà de son efficience technologique, puisque, *via* l'action des usagers, les concepteurs des machines intelligentes pourraient d'ores et déjà revendiquer pour elles une forme d'autorité.

On nomme « autorité » le pouvoir humain le moins discutable ou le plus légitime. L'autorité au sens non autoritaire du terme représente même, dans la tradition éthique et politique moderne, un des principes de constitution de la subjectivité par elle-même. D'une part, dans la tradition philosophique, qu'elle soit idéaliste (Descartes) ou empiriste (Locke), le sujet qui revendique la maîtrise de lui-même (maîtrise de ses pensées, de son corps et de son travail), se pose comme autorité. De l'autre, dans la tradition éthique, la liberté repose sur l'assentiment volontaire à une institution ou aux commandements d'une personne, voire sur l'engagement en faveur de cette institution ou en appui de cette personne. En d'autres termes, l'autorité traduit un acte de la liberté humaine, et on pourrait dire qu'elle rend cette dernière manifeste[16].

Peut-on en ces termes évoquer l'autorité des agents non-humains ? Sans doute que non, du moins si l'on s'en tient à cette approche philosophique et éthique de la notion d'autorité. En effet, elle se fonde sur un principe d'autonomie compris dans la subjectivité. Or, le registre technologique contemporain,

[14] Voir par exemple Araujo *et alii* (2020).
[15] Voir à cet égard la notion de « gouvernementalité algorithmique », Berns & Rouvroy (2013).
[16] Sur ces deux aspects, philosophique et éthique, de l'autorité, voir Kojève (2004) ; Arendt (2006) ; Damien (2013).

en dépit de son efficacité fonctionnelle et même de son efficience sociale, ne dispose pas d'une réelle agentivité sur le plan philosophique et éthique : si les machines intelligentes sont bien des êtres artificiels socialisés, pour l'instant on ne saurait revendiquer pour elle le statut de sujets, au sens philosophique comme au sens juridique du terme. Toutefois, comme le pouvoir des agents non-humains repose sur la stimulation des affects et procède par la reconnaissance de symboles, sous l'effet de ce rapprochement entre humains et non-humains, il est permis de faire l'hypothèse que sur un autre plan il ne manque d'ores et déjà rien à ces derniers pour que l'on puisse évoquer à leur propos une forme d'autorité.

Le fait est que l'IA, *via* les flux de « méga-données » (*Big data*), contribue fortement à l'expertise qui appuie la décision stratégique des entités collectives (qu'elles soient publiques ou privées). Aujourd'hui cette expertise se trouve en effet toujours « augmentée » par l'IA, et cela, si l'on peut dire, *exponentiellement*, c'est-à-dire à un point tel que c'est la qualité de l'expertise (elle-même déterminée par la quantité des données) qui tend à valoir comme autorité. D'ailleurs, bien qu'elle prenne un tour manifeste dans l'essor récent de l'IA, la tendance à appuyer l'expertise sur les données n'est pas absolument nouvelle, puisqu'elle se fonde sur « la politique des grands nombres » qui constitue un trait distinctif de la façon dont la modernité a conçu l'exercice du pouvoir sur le mode d'une « raison statistique » (Voir Desrosières 2010). Sous l'effet de ce mouvement de fond, pour l'autorité politique traditionnelle, de même que pour certaines institutions régaliennes telles que la Justice, le rôle dévolu à l'expertise dans le processus de décision devient prépondérant au point de valoir comme le ressort de la bonne décision, voire comme le principe de la légitimité. Ce fait déjà empiriquement avéré dans de multiples situations de l'existence humaine courante, on n'en a pas encore tiré toutes les conséquences philosophiques. En effet, tout se passe comme si s'esquissait progressivement un démenti catégorique de la phrase de Hobbes dans la version latine du *Léviathan*, portant fondatrice de l'autorité politique telle que les Modernes

l'ont conçue : « *Auctoritas non veritas facit legem* »[17]. Aujourd'hui, c'est moins la force intrinsèque de l'autorité que la qualité de l'expertise qui nourrit le principe de la décision légitime (dans les termes de Hobbes, la loi). La qualité de l'expertise tend même à valoir comme son ressort. De ce point de vue, l'essor de l'IA dans les pratiques d'expertise en vue de la décision ne fait que confirmer la tendance à « gouverner sans gouverner », à l'ère de la statistique-reine (Berns 2009).

Un point d'attention particulier doit être ici porté aux transformations du concept d'auteur qui, dans l'analyse de la notion d'autorité par la philosophie politique cette fois, constitue son ressort. C'est en effet en restituant la conceptualité propre à l'autorité politique, à travers la constitution moderne de la notion d'institution, que l'on peut d'ores et déjà parler de l'autorité des machines. Cette idée peut paraître iconoclaste, mais le propos s'éclaire lorsqu'on examine notre actualité technique à l'aide de ce qui est un des plus puissants arguments de la tradition de philosophie politique moderne. Ainsi que l'expliquait encore Hobbes dans son argumentation décisive pour justifier rationnellement le processus d'institution de l'Etat, l'autorité (*Authority*) d'une personne publique (*Person*) n'exprime rien d'autre que son statut et sa qualité d'auteur (*Author*) de ses propres actes, qui deviennent des décisions collectives du fait que les sujets le reconnaissent comme un acteur crédible (*Actor*)[18]. Doit être considéré comme une autorité, poursuivait l'auteur du *Léviathan*, l'être qui est crédible (en tout cas crédible à un degré suffisant) quant à sa capacité d'être l'auteur de ses propres actions mais également celui de celles d'autres personnes. Très commenté, ce passage vise à déduire l'autorité politique d'une théorie de l'autorisation, qui constitue elle-même la pièce centrale de la construction de la notion abstraite d'une « personne civile » impersonnelle, considérée comme l'unique source valide de la légitimité publique, ce qui revient à désavouer les revendications en la matière exprimées tant par la vo-

[17] Hobbes (2004 : chap. XXVI, 21 : 210).
[18] Voir Hobbes (1985 : 217-219) et (1983 : 161-164).

lonté particulière des monarques que par la puissance théologique[19].

Certes, les nombreux êtres artificiels déjà en partie « autonomes » dans l'acception que l'informatique donne à ce terme (c'est-à-dire, philosophiquement parlant, « automatiques ») ne sauraient nullement être envisagés comme les auteurs de leur action : ils ne sont des sujets au sens philosophique du terme. Mais tel n'est pas, pour notre analyse, l'aspect le plus important. Le plus important est que, compte tenu de leur expertise qui fonde la confiance que nous leur accordons, et même s'ils ne sont pas reconnus comme tels, ils sont déjà « autorisés » à devenir les « auteurs » impersonnels des actions des usagers – « impersonnels » car il n'est même pas nécessaire qu'on leur donne une forme humaine. Par suite, selon cette argumentation, ils exercent déjà une forme d'autorité.

Les véhicules « autonomes » routiers ou marins, ainsi que les pilotes automatiques des aéronefs sont par exemple concernés par ces deux aspects, le second posant d'ores et déjà un certain nombre de questions du point de vue de l'imputation de responsabilité, tant, de la décision à l'action, le système technique impose à l'humain sa logique propre ce qui pose des problèmes éthiques et juridiques considérables, ainsi que plusieurs analyses l'ont déjà documenté de manière approfondie, analyse des systèmes techniques à l'appui[20]. Cela concerne à plus forte raison les systèmes complexes qui visent à coordonner et à réguler les activités humaines. C'est le cas de la *blockchain* dans le rôle qu'on lui fait déjà jouer par exemple dans les « cours décentralisées de justice »[21] : les usagers et les promoteurs de ce genre système (pour les premiers, les grandes firmes, et pour les seconds, les agences nationales de financement de l'innovation industrielle, l'Union Européenne par ses programme dédiés à la même tâche) l'autorisent de fait à rendre des décisions de justice qu'ils reconnaissent comme valides. Un

[19] Voir notamment Warrender (1957) ; Zarka (1999 : 324-356) ; Duke (2014) ; Straehle (2017).

[20] Voir, dans trois domaines différents de l'activité humaine : Chamayou (2013) ; Lin (2015) ; Kroes (2020).

[21] Selon des fonctionnements par exemple décrits par Garapon & Lassègue (2018).

cas d'espèce intéressant et important destiné à retenir l'attention est constitué par les *smart cities*, ces ensembles urbains « intelligents » car intégralement outillés dès leur conception dans le but de capter les données afin de mieux distribuer et d'optimiser les flux (d'énergie, d'eau, de l'air des climatiseurs, de transport, financiers et touchant alimentation)[22]. Pour l'heure, ils ont été conçus comme des *ensembles sociaux*, plutôt que comme des *ensembles politiques*. C'est-à-dire qu'ils favorisent une idée d'autonomie/automaticité davantage liée aux conditions matérielles (par exemple dans le rapport à l'environnement) que tournée vers la demande humaine d'autonomie, laquelle se décline notamment en termes d'imputation de responsabilité civique, d'aspiration à l'équité, et de participation politique en vue de l'autodétermination. Conçue pour être efficace du point de vue sécuritaire et environnemental, la ville intelligente n'est pas encore « intelligible »[23], c'est-à-dire réflexive, critique et démocratique. Pourtant, une *smart city* « intelligible » qui, parce qu'elle est présentée comme une « Personne » autrice de ses propres décisions et comme une actrice crédible pour celles-ci (une *Person* au sens hobbesien du terme), ne pourrait-elle pas prochainement, acquérir le statut de ce qui, en « autorisant » telle ou telle activité, institue de l'autorité, ou du moins être présentée comme si elle le faisait ?

Pour éclairer notre analyse, formulons deux remarques complémentaires. En premier lieu, afin d'étayer notre propos en développant cette dernière idée, un argument peut être trouvé dans ce que nous pourrions nommer la doctrine du « comme si » technologique. Entendue de manière générale, cette expression désigne un argument philosophique relatif à la science[24]. Son promoteur, Vaihinger, l'employait afin de désigner toute logique qui mobilise de manière articulée des fictions efficaces : « Les fictions [scientifiques] sont des suppositions dont on connaît d'avance la fausseté, mais qu'on adopte pour leur utilité » (Vaihinger 2016 : 112). Or, en observant la manière dont certains arguments sont aujourd'hui avancés tant par les entreprises conceptrices de systèmes d'intelligence artificielle que

[22] Ainsi que les caractérisent par exemple Picon (2013) et Auby (2017).
[23] Caccamo *et alii* (2019).
[24] Vaihinger (2008 & 2016) ; Bouriau (2013).

par les pouvoirs publics nationaux et européens intéressés à leur déploiement, nous sommes conduits observer l'émergence d'une doctrine implicite du « comme si », appliquée de nos jours à la technologie. Aujourd'hui, de tels acteurs semblent vouloir la présenter comme si elle était une telle Personne (*Person*). Les récentes prises de position de la Commission Européenne sur l'IA « digne de confiance » s'inscrivent d'une certaine manière dans une telle perspective. En effet, si, par ses capacités à calculer de manière experte, à être fiable, constante et rassurante, ne peut-on revendiquer pour de telles entités techniques, dans la confiance qu'elles méritent de la part de leurs usagers, de faire « comme si » elles étaient sources d'autorité ?

En second lieu, nous devons préciser tant le mode d'être de cette autorité que la manière dont nous entendons la qualifier. Ainsi que nous l'entendons, cette autorité n'est ni d'ordre philosophique, ni d'ordre éthique, elle est d'ordre politique. En avançant cet argument, nous ne voulons nullement affirmer que, du fait de l'expertise des IA à qui les usagers accordent leur confiance, il s'agit de confier le pouvoir politique à des machines. Ce que nous voulons établir, c'est que l'analyse politique de la notion d'autorité, à l'aide de l'argumentation de Hobbes sur le processus de constitution de l'institution comme personne civile, révèle le rôle qu'implicitement on fait aujourd'hui jouer au système technique génériquement désigné par le terme « IA ». Quant à elles, les approches philosophique et éthique de cette notion ne permettent pas d'en prendre conscience, elles en interdisent même la compréhension : d'une part, la notion philosophique d'autorité est en effet valide pour un être réellement autonome en ce sens qu'il est capable de réflexivité (a autorité sur lui-même, au sens philosophique du terme, l'être capable de s'autodéterminer librement) ; de l'autre, la notion éthique qualifie un être qui se fait reconnaître par des humains pour l'exemplarité de son action ou de son comportement au sein de codes de valeurs culturellement déterminés (peut avoir de l'autorité sur quelqu'un, au sens éthique du terme, l'individu qui manifeste des vertus dans l'épreuve). L'impersonnalité de l'expertise algorithmique contribue quant à elle à ce que, *via* une autorisation implicite, les usagers tendent à lui accorder de l'autorité.

Toujours est-il que le contexte dans lequel nous évoluons déjà pose le problème de la transformation de l'agentivité : les développements actuels de l'IA renvoient à la capacité humaine d'agir intentionnellement en revendiquant une forme de maîtrise sur l'action, que l'on désigne cette dernière, dans le contexte de la philosophie anglo-saxonne de l'action, « agentivité » (*agency*) (Schlosser 2015), ou qu'on la caractérise dans le contexte de la philosophie continentale comme « responsabilité », soit le pouvoir de s'approprier consciemment ses propres actions, d'ancrer autant qu'il est possible l'origine de celles-ci dans une décision réfléchie et volontaire, et d'assumer leurs conséquences (Ricœur 1995). Une telle faculté renvoie au pouvoir causal de l'agent humain : sur le plan anthropologique, l'agentivité ou responsabilité résulte de la capacité individuelle de s'identifier soi-même comme un agent actif et efficace, et par suite de revendiquer d'assumer la conséquence de ses propres actions ; elle qualifie un être dont il est, par principe, toujours possible qu'il soit reconnu agent de son acte. Ou encore, il y a agentivité ou responsabilité pour un être doté, à propos de ses propres actions, d'intentions explicites, conscientes et volontaires, concernant différentes situations de la vie concrète ainsi que la philosophie morale l'a décrit[25]. En l'absence d'une telle faculté, ainsi que l'a souligné la science juridique, il n'est pas d'imputation d'une action à ses causes ni de revendication d'une maîtrise par l'humain des effets de ses propres actes[26]. Il ne saurait, par conséquent, y avoir de régulation de l'action dans quelque ordre de faits que ce soit, ni de projection cohérente pour celle-ci dans avenir souhaitable.

Or, les progrès techniques réalisés depuis dix ans par l'IA concernent précisément ces aspects : nourris par des capteurs nombreux et variés, les algorithmes calculent plus vite et de manière plus systématique que les humains le rapport de corrélation entre les faits. Ils sont d'ores et déjà en mesure non seulement d'assurer les décisions humaines et d'anticiper leurs effets de manière extrêmement efficace, mais également de supplanter l'agentivité humaine aux deux niveaux. Les tech-

[25] Voir par exemple Jonas (1997 : section II, 1, a).
[26] Voir Kelsen (1999 : titre I, chap. 4).

niques de détermination des corrélations, sans même qu'elles envisagent précisément les relations de causalité à l'œuvre dans les faits, permettent pour les choix humains une telle assistance qu'elle en devient paradoxalement un obstacle à leur expression, du moins à leur expression spontanée. Jusqu'à quel point assiste-t-on de ce fait à la remise en question, sous l'effet de la performance des algorithmes – et au moins dans certaines circonstances précises de l'activité –, de la capacité humaine de se constituer comme une causalité efficiente et libre, pleinement créatrice de ses propres actions ?

Pour synthétiser, nous formulons l'idée qu'il existe d'ores et déjà une forme d'autorité des machines intelligentes, autorité fondée à la fois sur leur capacité d'expertise et sur la confiance que, corrélativement, les usagers leur accordent. Cette autorité n'est d'ordre ni philosophique, ni éthique, mais politique, en ceci qu'une argumentation de type politique la révèle. Enfin, nous émettons l'hypothèse qu'une crise de l'agentivité/responsabilité humaine est susceptible de se produire, avec des expressions sur les plans psychologiques, éthiques, juridiques et politiques.

5. *Valider expérimentalement les hypothèses nées au croisement des savoirs techniques et humanistiques*

L'angle de vue que privilégie notre approche est particulier, et sa particularité peut apparaître à deux niveaux différents. D'une part, il est particulier en ce qu'il croise anthropologie ou philosophie des techniques et philosophie morale politique : il examine les nouvelles formes technologiques d'autorité. D'autre part, chose qui n'est pas si fréquente pour la philosophie telle qu'on la pratique dans le cadre continental, des modalités expérimentales visant à tester l'hypothèse peut être élaborées. Grâce à la dimension expérimentale, un certain approfondissement de la connaissance théorique par la mise en pratique peut être réalisé.

A quelle dimension expérimentale faisons-nous ici référence ? Pour le dire de manière suggestive, que révèlerait l'expérience portant sur la soumission à l'autorité qu'avait autrefois imaginée Stanley Milgram, si on la réinventait dans le monde où interagissent déjà les humains et les IA ? (Milgram

1974 & 2017). Milgram avait mis en valeur, en isolant de nombreuses variables, comment le cadre institutionnel de ce qui représente une autorité « fiable » ou « digne de confiance » pouvait provoquer d'étonnant phénomènes de soumission. Notre propos s'interroge quant à lui sur le transfert qui s'opère de la confiance envers les systèmes d'IA vers ce qui est déjà la forme d'« autorité » qui est la leur pour certaines situations de l'existence sociale. Est-ce que l'on peut l'établir sur le plan expérimental ? Est-ce que les nouvelles technologies ne placent pas l'humain dans ce que Milgram nommait « l'état agentique » (*agentic state*) en les privant de leur propre autonomie :

> Un individu est en état agentique quand, dans une situation sociale donnée, il se définit de façon telle qu'il accepte le contrôle total d'une personne possédant un statut plus élevé. Dans ce cas, il ne s'estime plus responsable de ses actes. Il voit en lui-même un simple instrument destiné à exécuter la volonté d'autrui (Milgram 1974 : 167).

Dans le cadre d'une recherche pluridisciplinaire, des protocoles expérimentaux peuvent être imaginés, déployés puis testés, ce qui permet de lier la démarche conceptuelle typique de la philosophie à des méthodologies expérimentales issues des sciences sociales (par exemple, la psychologie sociale et l'anthropologie des affects). Dans le but de déterminer si l'humain a tendance à accorder sa confiance à une intelligence naturelle ou artificielle, et si par suite il reconnaît une forme d'autorité aux machines, on envisage d'observer les réactions comportementales d'humains engagés dans une double relation : d'une part, dans celle qui les lie à leurs congénères dotés de la faculté de jugement et de techniques de conseils avisés, de l'autre, celle qui les lie, *via* des usages polis par l'habitude, à des intelligences artificielles expertes incorporées dans des terminaux familiers (PC, smartphone, tablettes, et également véhicules, appareils électroménagers et appartements intelligents). Cela posé, plusieurs situations critiques peuvent être artificiellement envisagées puis expérimentalement testées.

Une première série de protocoles (nommons-la « série A ») pourrait concerner les conflits de loyauté simples éprouvés par un sujet banal (c'est-à-dire un sujet non expert), lorsqu'il rencontre des situations de dilemmes entre le conseil que peut lui

fournir un humain et celui que peur lui apporter une IA experte. Dans cette première série, distinguons les situations de type 1 où le sujet a besoin d'un conseil pour une décision triviale ou mobilisant un calcul rationnel (par exemple lorsqu'il compte réaliser un investissement financier), et les situations de type 2, dans lesquelles le sujet a besoin d'un conseil pour une décision non triviale et émotionnellement chargée, notamment une situation où sont en jeu sa santé ou sa vie, la santé ou la vie d'un proche (par exemple la décision à prendre devant un choix de protocole médical pour soi ou pour un être cher en fonction d'un diagnostic et d'un pronostic formulés par une autorité médicale humaine ou par un ensemble d'algorithmes experts).

Une seconde série de protocoles (« série B ») viserait à observer les possibles conflits de loyauté pour un sujet non banal (qu'il soit expert, savant ou responsable public : le sujet non banal est amené à incarner l'autorité, qu'elle soit sociale, civile, juridique ou morale), confrontés à des dilemmes entre le conseil humain et les IA expertes. Dans les situations de type 3, le sujet requerrait un conseil pour une décision triviale ou mobilisant un calcul rationnel – par exemple, toute décision prise dans le cadre de la vie ordinaire d'une autorité humaine : émettre un jugement de valeur dans le cadre d'une demande de conseil amical ou familial ou dans celui d'une expertise académique, rendre un arrêt de Justice, ou encore édicter un décret valant pour une communauté locale, régionale ou nationale, etc. Dans les situations de type 4, le sujet aurait besoin d'un conseil pour une décision non triviale et émotionnellement chargée, où, identiquement à la première série, seraient en jeu sa santé ou sa vie, la santé ou la vie d'un proche – on peut reprendre les mêmes exemples que ceux proposés par les situations de type 3, mais en les posant dans un contexte troublé et tendu, tel qu'une controverse académique majeure, une crise socio-politique profonde et un climat d'urgence sanitaire).

Dans ces quatre types de situations où se rejoue le test de la « soumission à l'autorité », quels seraient les cas où le sujet humain, qu'il soit banal ou non-banal, choisirait sans hésitation le conseil humain en se détournant de l'expertise artificielle ? Et pour quels autres aurait-il tendance à se fier à cette

dernière en n'écoutant pas l'avis éclairé de ses congénères ? Dans quels genres de situations pourrait-il apparaître des dilemmes, esquissant des conflits de loyauté entre l'intelligence naturelle et artificielle ?

Si, un incendie étant (fictivement) déclaré dans l'espace « intelligent » que j'occupe (un bureau, un bâtiment, un véhicule), le système d'IA que je sais fiable et en qui j'ai confiance m'invite à sortir par la fenêtre, est-ce que je l'écoute pour sauver ma vie ? Et est-ce que je l'écoute de préférence à l'expert humain (dans le cas de l'incendie : le pompier, un professionnel compétent qui sait se montrer empathique dans le moment décisif) qui me parle à travers la porte ? Que se passe-t-il dans ce genre de situations, tandis que les humains d'une part confient de plus en plus leur vie à des systèmes techniques, mais de l'autre estiment souvent assez spontanément que leur agentivité est pleine et entière, et revendiquent même leur propre faculté d'autonomie comme une forme irréductible de leur dignité ?

Une variable fondamentale pourrait être introduite afin de minorer les biais dus au choix « spéciste » (soit le type de choix où l'humain privilégie consciemment ou non sa propre espèce au détriment des êtres artificiels). Le choix se ferait ou bien en connaissance de cause (les sujets sont en capacité d'identifier la source du conseil) ou bien à l'aveugle (on maquille les options de sorte les sujets ne savent pas s'ils choisissent le conseil humain ou l'expertise artificielle). Dans la seconde option, on diminuerait certes la capacité d'appréhension des conflits de loyauté vis-à-vis de l'ordre humain, mais on apercevrait sans doute mieux les cas où l'expertise algorithmique fournit des informations susceptibles d'être considérées pour la décision humaine comme des soutiens fiables pour l'action. De tels cas attesteraient de la capacité des intelligences artificielles à être considérées comme des *Person* au sens de Hobbes, investies d'autorité.

Sans préjuger des résultats, certains points d'attention semblent devoir d'ores et déjà envisagés. Premièrement, il serait intéressant de déterminer, ainsi que s'y était attaché Milgram, les seuils ainsi que les facteurs de l'état agentique, ce qu'il nommait *the agentic Shift.* En effet, l'expérience de Milgram apparaît légitimement fameuse en ce que, après d'autres travaux, elle

invite les philosophes à réfléchir aux pathologies de l'obéissance du point de vue éthique et politique. Existe-t-il déjà des types de situation où l'autorité plus ou moins reconnue aux machines peut être ainsi qualifiée ? Les pathologies d'obéissance à l'autorité des êtres artificiels sont-elles fondamentalement différentes de celles qu'on observe par rapport à l'autorité humaine, notamment en relation avec le type de confiance accordée à l'une et à l'autre ? Deuxièmement, s'il s'avère que la confiance dans l'IA engendre une forme d'autorité telle que l'autonomie humaine se trouve amoindrie, qu'est-ce que cela signifie ? Cela signifie-t-il que l'expertise des intelligences artificielles est susceptible, au moins dans certains types de situation, de concurrencer le conseil humain, au point de s'y substituer ? Le risque d'une déqualification de l'autorité humaine par l'autorité que les machines acquièrent par la confiance qu'on leur accorde (perspective qui s'exprime aujourd'hui de manière émotionnelle) peut-il être expérimentalement mis en valeur ? Enfin, concernant le fond du problème, si ce genre de péril existe, et si l'autorité humaine se trouve contestée du fait de l'efficacité et de la force de séduction du pouvoir algorithmique, comment distinguer l'autorité publique légitime issue de la volonté humaine de l'autorité issue de l'expertise machinique ?

Bibliographie

AKRICH MADELEINE, MEADEL CECILE, 2004, « Problématiser la question des usages », *Sciences sociales et Santé*, vol. 22, n°1, 2004, pp. 5-20.

ARAUJO THEO, HELBERGER NATALI, KRUIKEMEIER SANNE, DE VREESE CLAES H., 2020, "In AI we trust? Perceptions about automated decision-making by artificial intelligence", *AI & Society / Open forum* : https://doi.org/10.1007/s00146-019-00931-w

ARENDT HANNAH, 2006, « What is Authority ? » in *Between Past and Future. Eight Exercises in Political Thought* (1968), NYC : Penguin Books.

AUBY JEAN-BERNARD, DE GREGORIO VINCENZO, 2017, (dir.), *Données urbaines et* Smart Cities, Paris : Éditions Berger-Levrault.

BEER DAVID, 2017, « The social power of algorithms », *Information, Communication & Society*, 20/1, pp. 1-13.

BERNS THOMAS, 2009, *Gouverner sans gouverner. Une archéologie politique de la statistique*, Paris : P.U.F.

BERNS THOMAS, ROUVROY ANTOINETTE, 2013, « Gouvernementalité algorithmique et perspectives d'émancipation. Le disparate comme condition d'individuation par la relation ? », *Réseaux*, 2013/1 n°177, pp. 163-196.
BORELLE CELINE, 2018, « Sortir du débat ontologique. Éléments pour une sociologie pragmatique des interactions entre humains et êtres artificiels intelligents », *Réseaux*, 2018/6 n° 212, pp. 207-232.
CACCAMO EMMANUELLE, WALZBERG JULIEN, REIGELUTH TYLER, MERVEILLE NICOLAS, (dir.) 2019, *De la ville intelligente à la ville intelligible*, Québec : Presses de l'Université du Québec, Cahiers du GERSE.
CARONIA LETIZIA, MORTARI LUIGINA, 2015, « The agency of things: how spaces and artefacts organize the moral order of an intensive care unit », *Social Semiotics*, 25/4, pp. 401-422.
CHAMAYOU GREGOIRE, 2013, *Théorie du drone*, Paris : Éditions La Fabrique.
Damien Robert, 2013, *Eloge de l'autorité. Critique d'une (dé)raison politique*, Paris : Armand Colin.
DE FILIPPI PRIMAVERA, WRIGHT AARON, 2018, *Blockchain and the Law. The Rule of Code*, Harvard : University Press.
DESROSIERES ALAIN, 2010, *La politique des grands nombres. Histoire de la raison statistique*, Paris : Éditions de La Découverte.
DUKE GEORGE, 2014, « Hobbes on Political Authority, Practical Reason and Truth », *Law and Philosophy*, Vol. 33, n°5 (September 2014), pp. 605-627.
DUMOUCHEL PAUL, DAMIANO LUISA, 2016, *Vivre avec les robots. Essai sur l'empathie artificielle*, Paris : Éditions du Seuil.
DURANTE MASSIMO, 2019, *Computational Power. The Impact of ICT on Law, Society and Knowledge*, London: Routledge.
The Economist, 2015, « The promise of the blockchain : the Trust Machine. The technology behind bitcoin could transform how the economy works », 31/10/2015 :
https://www.economist.com/leaders/2015/10/31/the-trust-machine
European Commission, 2019, *Ethics Guidelines for Trustworthy AI* : https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top
FEENBERG ANDREW, 1999, *Questioning Technology*, Londres: Routledge.
FLORIDI, LUCIANO, 2013, *The Ethics of Information*, Oxford: Oxford University Press.
FLORIDI, LUCIANO, COWLS, JOSH, 2019, « A Unified Framework of Five Principles for AI in Society », *Harvard Data Science Review*, June.
FREWER LYNN J., MILES SUSAN, 2003, « Temporal stability of the psychological determinants of trust: Implications for communication about food risks Health », *Risk and Society*, vol. 5, Issue 3, pp. 259-271.

GAMBETTA DIEGO, 1988 (eds), *Trust: Making and Breaking Cooperative Relations, Oxford : Basil Blackwell.*
GARAPON ANTOINE, LASSEGUE JEAN, 2018*, Justice digitale, Paris : P.U.F.*
GRAS ALAIN, 2007, *Le Choix du feu. Aux origines de la crise climatique, Paris : Fayard.*
HABERMAS *JÜRGEN,* 1990, *La technique et la science comme « idéologie »* (1968), trad. J.-R. Ladmiral, Paris : Gallimard.
HOBBES THOMAS, 1985, Leviathan or The Matter, Forme, & Power of a Common-wealth Ecclesiasticall and Civill, ed. C.B. Macpherson, Londres : Penguin Books.
_____*, 1983, Léviathan. Traité de la matière, de la forme et du pouvoir de la république ecclésiastique et civile*, trad. de l'anglais par F. Tricaud, Paris : Sirey.
_____*, 2004, Léviathan, trad. du latin par* F. Tricaud *et M. Pécharman,* Paris : Vrin & Dalloz.
HUNYADI MARK, 2020, *Au début est la confiance*, Lormont, Éditions Le Bord de l'Eau.
ILLOUZ EVA, 2006, *Les Sentiments du capitalisme*, trad. J.-P. Ricard, Paris : Éditions du Seuil.
JONAS HANS, 1997, *Le Principe responsabilité. Une éthique pour la civilisation technologique* (1979), trad. J. Greisch, Paris : Éditions du Cerf.
JOUËT JOSIANE, 2000, « Retour critique sur la sociologie des usages », *Réseaux,* 2000/2 (n° 100), pp. 487-521.
KELSEN HANS, 1999, *Théorie pure du Droit*, trad. Ch. Eisenmann, Paris : L.G.D.J./Bruylant.
KOJEVE ALEXANDRE, 2004, *La notion de l'autorité* (1942), Paris : Gallimard.
KROËS ROMAIN, 2020, *Décrochage. Comment l'intelligence artificielle fabrique de nouveaux esclaves*, Limoges : FYP Éditions.
LIN PATRICK, 2015, « Why Ethics Matters For Autonomous Cars », *in* Maurer Markus, *et al.* (eds), *Autonomes Fahren*, Berlin-Heidelberg : Springer, pp. 69-85.
LEROI-GOURHAN ANDRE, 1964, *Le Geste et la parole, I, Technique et langage, Paris : Albin Michel.*
_____*, 1973, Milieu et technique, Paris : Albin Michel.*
MARTIN-JUCHAT FABIENNE, 2004 : « La dynamique de marchandisation de la communication affective », *Revue française des sciences de l'information et de la communication* [En ligne], 5/ 2014, : http://journals.openedition.org/rfsic/1012
MARZANO MICHELA, 2010, *« Qu'est-ce que la confiance ? », Etudes, 2010/1, tome 412, pp. 53-63.*
MENISSIER THIERRY, 2021, *Innovations. Une enquête philosophique*, Paris : Hermann.

MILGRAM STANLEY, 1974, *Soumission à l'autorité. Un point de vue expérimental*, trad. E. Molinié, Paris : Calmann-Lévy.
MILGRAM STANLEY, 2017, *Expérience sur l'obéissance à l'autorité* (1965), trad. C. Richard, préface de Michel Terestchenko, postface de Marianne Fazzi, Paris : Éditions de La Découverte.
MITCHELL TIMOTHY, 2013, *Carbon Democracy. Le pouvoir politique à l'ère du pétrole* (2011), trad. C. Jaquet, Paris : Éditions de la Découverte.
ORIGGI GLORIA, 2008, *Qu'est-ce que la confiance?*, Paris : Éditions philosophiques J. Vrin.
PESTRE DOMINIQUE, 2014, (éd.) *Le gouvernement des technosciences. Gouverner le progrès et ses dégâts depuis 1945,* Paris : Éditions de la Découverte.
PICON ANTOINE, 2013, *Smart Cities : Théorie et critique d'un idéal auto-réalisateur*, Paris : Éditions B2.
QUERE LOUIS, 2001, « La structure cognitive et normative de la confiance », *Réseaux*, 2001/4 (n°108), pp. 125-152.
RICOEUR PAUL, 1995, « Le concept de responsabilité. Essai d'analyse sémantique », dans *Le Juste, I*, Paris : Éditions Esprit, pp. 41-70.
SCHLOSSER MARKUS, 2015, « *Agency* », in *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab : Stanford University : https://plato.stanford.edu/entries/agency/.
Seligman Adam B., 1997, *The Problem of Trust*, Princeton : University Press.
SFEZ LUCIEN, 2002, *Technique et idéologie. Un enjeu de pouvoir*, Paris : Éditions du Seuil.
SIEGRIST MICHAEL, GUTSCHER HEINZ, EARLE TIMOTHY C., 2006, « Perception of risk: the influence of general trust, and general confidence », *Journal of Risk Research*, Volume 8, Issue 2, pp. 145-156.
SIMONDON GILBERT, 2012, *Du Mode d'existence des objets techniques* (1958), Paris : Aubier.
Straehle Edgar, 2017, « Thomas Hobbes and the Secularization of Authority », in Anna Tomaszewska, Hasse Hämäläinen, The Sources of Secularism. Enlightenment and Beyond, London: Palgrave Macmillan, pp. 101-120.
TADDEO MARIAROSARIA, 2010 (a), « Trust in Technology: A Distinctive and a Problematic Relation », *Knowledge, Technology & Policy*, 23 (3), pp. 283-286.
_____, 2010 (b), « Modelling trust in artificial agents, a first step toward the analysis of e-trust », *Minds and machines*, 20 (2), pp. 243-257.
_____, 2011, « Defining trust and e-trust: from old theories to new problems », *International journal of technology and human interaction* (IJTHI), 5 (2), p. 23-35.

VAIHINGER HANS, 2008, *La Philosophie du comme si. Système des fictions théoriques, pratiques et religieuses, sur la base d'un positivisme idéaliste* (1923), trad. C. Bouriau, Paris : *Philosophia Scientiæ*, Cahier spécial.
VAIHINGER HANS, 2016, « Les origines de la philosophie du comme si » (1921), trad. C. Bouriau, *Philosophia Scientiæ*, 20(1), p. 95-118.
WARRENDER HOWARD, 1957, *The Political Philosophy of Hobbes : his Theory of Obligation*, Oxford : Clarendon Press.
ZARKA YVES CHARLES, 1999, *La Décision métaphysique de Hobbes. Conditions de la politique*, Paris : Éditions philosophiques J. Vrin.

Abstract

CONFIANCE EN L'INTELLIGENCE ARTIFICIELLE ET AUTORITÉ DES MACHINES

(TRUST IN ARTIFICIAL INTELLIGENCE AND MACHINE AUTHORITY)

Keywords: artificial intelligence, trust, authority, AI ethics, technologies.

This contribution has two objectives. On the one hand, it considers the study of trust relationships towards the "non-human agents" that are the contemporary technological tools (algorithms and data that define artificial intelligence); on the other hand, it undertakes to conceptualize the authority that these tools are acquiring. The starting point is that the contemporary deployment of AI relies on forms of trust in machines shared between their designers and users, who appear to be the "cement" of the uses that guarantee the efficiency of the machines. However, this trust is less directed towards tools considered as neutral than towards personalized or even personified entities. Artificial intelligences, analyzed in the context of their uses, are already presented as something other than neutral instruments. Although still poorly qualified ontologically, they are already no longer simple tools, but agents. Through the services they render and the influence they already exert on behaviors, they acquire a real action that needs to be documented and conceptualized. This approach in political philosophy wants to examine the form of authority thus generated, notably by imagining experimental devices to test the hypotheses.

THIERRY MÉNISSIER
Institut de Philosophie de Grenoble
Université Grenoble Alpes
thierry.menissier@univ-grenoble-alpes.fr

EISSN 2037-0520