Enjeux éthiques situés de l'IA

Manuel Zacklad
Dicen-IdF, CNAM
Manuel.zacklad@lecnam.net

Antoinette Rouvroy
FRS-FNRS, CRIDS, Université de Naumur
antoinette.rouvroy@fundp.ac.be

Mots-clefs : Ethique, Ethique située, Intelligence Artificielle, Controverse

Keywords: Ethic, Situated ethic, Artificial Intelligence, Controversy

Résumé

Pour aborder la manière dont l'IA affecte le quotidien, nous nous positionnerons dans une approche pragmatique de l'éthique, que nous appellerons éthique située, qui nous semble une alternative à des approches incantatoires courant de « l'IA éthique » qui suscite un certain nombre de critiques pointant un risque « d'éthique-washing ». Après une présentation des principes de l'éthique située qui ne sépare pas « royaume des valeurs » du « monde des faits » nous rappellerons les enjeux particuliers de l'IA connexionniste et la diversité de ses applications dans des situations quotidiennes.

Abstract

To address the way in which AI affects everyday life, we will position ourselves in a pragmatic approach to ethics, which we will call situated ethics, which seems to us an alternative to the incantatory approaches common to "ethical AI" which arouses a a number of critics pointing to a risk of "ethics-washing". After a presentation of the principles of situated ethics which does not separate the "realm of values" from the "world of facts", we will recall the particular issues of connectionist AI and the diversity of its applications in everyday situations.

Enjeux éthiques situés de l'IA

Manuel Zacklad
Antoinette Rouvroy

Introduction

Dans cette communication1, nous aborderons les enjeux éthiques liés à l'automatisation des dispositifs d'information et de communication qui affectent les comportements des personnes dans leur vie quotidienne : sélection-hiérarchisation automatique de contenus, accès aux services bancaires, aux assurances, à l'emploi, à la justice, à l'orientation scolaire, recommandation d'information, ou de produits de loisir, etc… En particulier nous nous intéresserons à la conception et à l'exploitation des algorithmes de l'Intelligence Artificielle connexionniste qui relèvent notamment de l'apprentissage profond (deep learning) et qui ont permis de très nombreuses innovations pratiques ces dernières années : traduction automatique, reconnaissance d'image, recommandation, « décision » automatique, etc.

Nous nous positionnerons dans une approche pragmatique de l'éthique, que nous appellerons éthique située, qui nous semble une alternative pertinente à des approches parfois un peu incantatoires de la relation entre éthique et IA (cf. le rapport Villani 2017) et en particulier au courant de « l'IA éthique » qui suscite un certain nombre de critiques pointant un risque « d'éthique-washing » (Ochigame 2019) mais aussi au courant de recherche de l'éthique de la technologie (Wright 2011). Après une présentation de l'éthique située nous rappellerons les enjeux particuliers de l'IA connexionniste et la diversité de ses applications dans des situations quotidiennes[2][3].

---

[1] Cette présentation prolonge une présentation orale donnée à Séoul par le premier auteur dans la conférence institutionnelle « AI for trust - 1st International Conference on Ethics of the Intelligence Information Society », Séoul, 5 décembre 2019 et est une version raccourcie d'un texte original qui sera soumis à la RFSIC à l'occasion de l'appel « Questionner l'éthique depuis les SIC en contexte numérique ».

[2] Les dimensions éthiques ne sont pas abordées ici du point de vue de l'activité des chercheurs comme le font certains travaux en SIC (Domenget et Wilhelm 2017)

[3] Dans la version longue nous introduisons les six espaces de controverses de l'éthique située de l'IA connexionniste qui pour certains concernent les applications numériques dans leur globalité.

L’éthique située

Comme le rappellent Lobet-Maris et ses co-auteurs (Lobet-Maris et al. 2019) « une nouvelle contrainte a été inscrite dans les politiques scientifiques internationales et s’est retrouvée in extenso dans les projets de recherche & développement (r&d) européens [notamment dans le domaine de la sécurité]. Il s’agit de la nécessaire prise en compte des enjeux éthiques, juridiques et sociaux au cœur de l’innovation technologique » désignée par l’acronyme RRI (Responsable Research Innovation). Ces efforts pour introduire l’éthique dans les projets technologiques et notamment biotechnologiques s’inscrivent dans la suite de travaux initiés depuis de début des années 2000 qui sont bien représentés par la proposition de « cadrage éthique » de Wright (2011).

Pour Wright, qui s’inscrit lui-même dans la continuité des travaux de l’éthique biomédicale de Beauchamp et Childress (2011), l’évaluation éthique des technologies de l’information vise essentiellement à étudier leur impact sur les utilisateurs ou la société. Il reprend à ces auteurs les dimensions du respect de l’autonomie (consentement éclairé), de l’abstention de nuire (non-maléficience, sureté, isolement et privation du contact humain, discrimination…), bénéficience (orientation de l’action vers le bien, proportionalité des moyens et des fins) et de la justice (accessibilité, solidarité sociale, inclusion et exclusion, non-discrimination, égalité d’opportunités, solidarité sociale, inclusion). Il rajoute, pour prendre en compte les spécificités du numérique, les dimensions de la vie privé et de la protection des données qui inclut les questions de qualité des données, de limitation de l’utilisation des données, de transparence au sens de l’ouverture des données, de l’accès des individus à leurs données et de leur participation à leur mise à jour, d’anonymat, de respect des communications privées (traçabilité) et enfin de respect du caractère privé du comportement personnel. Dans un contexte européen ces dimensions sont déjà inscrites dans la Convention européenne des droits de l’Homme, dans la Charte des droits fondamentaux de l’Union européenne et dans le RGPD.

Les outils éthiques déployés relèvent des études d’impacts classiques à base de consultation et de sondage, d’atelier d’expert, de liste de questions, d’une matrice éthique, d’un Delphi éthique auquel il rajoute l’idée de conférence de consensus et de panel citoyen. Cette approche essentiellement « externe » consiste à considérer le projet technologique comme une donnée dont il faut analyser les impacts sur la population en s’appuyant sur un certain nombre de valeurs a priori.

Par contraste, l'éthique située telle que nous la définissons, s'appuie à la fois sur le pragmatisme de J. Dewey et sur la philosophie des sciences. Dans la théorie de la valuation (Dewey 2008/1939, Prairat 2014) Dewey s'oppose à deux approches de la valeur, celle qui en ferait l'expression de préférences émotionnelles et passagères des acteurs et celles qui, à l'inverse, en ferait des fins en soi s'imposant aux acteurs de l'extérieur de manière quasi transcendante. Les valeurs sont pour Dewey « un produit de l'activité intelligente ouvert à l'éducation du regard et du jugement » qui résultent de l'expérience.

Pour lui, les valeurs sont en fait la cristallisation de processus de valuation qui comportent deux temps. Une première appréciation subjective, la valorisation, basée sur le désir et la prise en compte de sa satisfaction, suivie par un deuxième temps d'évaluation qui met en perspective les moyens consentis et les avantages procurés. Avec le temps, les règles issues du processus de valuation deviennent des normes qui guident l'action et qui peuvent avoir tendance à s'autonomiser, à devenir abstraites, si les sujets oublient les expériences qui avaient conduit à leur installation. Mais dans le fond, ces normes sont justifiées par sur une assertablité garantie par des expériences répétées. Chez Dewey « on ne saurait donc séparer de manière étanche le « royaume des valeurs » du « monde des faits », une position que l'on retrouve chez les philosophes des sciences, comme chez E. Hache (2011) qui, dans ses réflexions pour une éthique environnementale, en appelle également à une éthique qui ne sépare pas la question des faits et celle des valeurs ou la science de la morale et à considérer qu'il y a une objectivité des valeurs comme il y a une objectivité des faits.

C'est la principale différence entre l'éthique de la technologie de Wright, dont se revendique également Loblet-Maris et ses collaborateurs (2019) et l'éthique située. Les technologies ne sont pas des faits inéluctables dont il faut étudier les impacts en faisant appel à des principes moraux qui s'imposent de l'extérieur. Au contraire, l'émergence d'un problème éthique suscité par une nouvelle technologie invite à remettre en cause ses présupposés scientifiques, techniques, économiques, etc. considérés comme acquis, pour les examiner sous un angle pluridisciplinaire et orienter les développements technologiques de manière différente. La réorientation ne vient pas d'une contrainte morale externe, elle est issue d'éclairages scientifiques et politiques nouveaux qui remettent en cause certaines croyances et suggèrent d'autres pistes de recherche et de développement.

Par exemple, les biais des algorithmes connexionnistes qui discriminent les populations minoritaires, sont considérés comme des problèmes éthiques parce qu'il y a une contradiction entre la croyance quant à la supériorité des données « objectives » et de l'IA qui les exploite

pour fournir des décisions automatiques, d'un côté, et les erreurs constatées, de l'autre[4]. L'approche éthique externe qui vise à atténuer les impacts consiste, d'une part, à chercher à encadrer d'un point de vue juridique la décision automatique et d'autre part, à chercher à améliorer l'explicabilité des algorithmes pour les rendre plus compréhensibles (Besse et al. 2019).

L'éthique située va consister à remettre en cause les présupposés scientifiques des promoteurs de la décision automatique en contestant la possibilité qu'il existe des données objectives et en montrant le caractère intrinsèquement opaque de l'apprentissage profond, ce qui interdit radicalement des explications « logiques », comme nous allons le développer plus bas. Cette remise en cause n'est pas le fait d'éthiciens professionnels qui seraient les garants de valeurs transcendantes mais résulte de la capacité à poser autrement les problèmes de nature, scientifique, sociale, politique en ouvrant des espaces de controverses pluridisciplinaires. Cette approche est bien sûr plus difficile à mettre en œuvre dans les projets financés par les gouvernements qui s'inscrive souvent une logique « taylorienne » (cf. Lobet-Maris et al. 2019).

Mais l'approche éthique externe souffre aussi d'une autre difficulté. Nous avons vu qu'elle sépare la question des valeurs de celle des faits en tentant de réduire les impacts des changements technologiques sans remettre en cause leurs présupposés ou en cherchant à atténuer à la marge leurs nuisances par des correctifs techniques mineurs. Mais elle souffre aussi du fait d'envisager la valeur sans prendre en compte les questions de participation dans la définition même des problèmes à traiter. En effet, l'émergence du problème éthique et des conflits de valeur associés ne saurait être simplement circonscrits par des experts ou des sondages.

Dans la perspective de l'éthique située, il faut également objectiver les intérêts des acteurs qui sont des parties prenantes identifiées des « solutions » envisagées mais aussi parfois des parties prenantes indirectes qui n'ont pas été prises en compte par les promoteurs du projet. Si l'objectivation des contre-arguments factuels passe par la mise en place de controverses scientifiques pluridisciplinaires, elle doit s'accompagner par l'objectivation du conflit de valeur qui est de nature politique et qui doit s'incarner par la constitution d'un public au sens

[4] Dans le domaine de l'IA connexionniste, la notion d'erreur pourrait elle-même être discutée dans la mesure ou les « décisions » erronées reflètent l'état des pratiques sociales, comme la surreprésentation des noirs dans le système carcéral américain (logiciel de prévention sécuritaire) ou la place avantageuse des hommes dans la hiérarchie des entreprises (logiciel de recrutement), cf. infra, controverse liée aux enjeux de culture numérique.

de J. Dewey dans son ouvrage « Le public et ses problèmes » (1927/2010). Comme le rappelle Joelle Zask (2008) :

« *Un public est l'ensemble des gens ayant un plein accès aux données concernant les affaires qui les concernent, formant des jugements communs quant à la conduite à tenir sur la base de ces données et jouissant de la possibilité de manifester ouvertement ses jugements. On doit lui reconnaître une autorité en la matière, un droit d'exercer son jugement et une grande liberté dans le choix des moyens nécessaires à le faire entendre : opinion publique, presse, Internet, associations, débats publics et ainsi de suite. L'autorité du public suppose donc une liberté d'enquête, une pleine information, une éducation appropriée pour acquérir la compétence d'évaluer les corpus documentaires, voire de les constituer, et des droits politiques garantis. L'ensemble de ces conditions est décliné dans Le public et ses problèmes*. »

Pour Dewey c'est le fait de se sentir concerné par un problème commun et de souhaiter se mobiliser pour en trouver la solution qui constitue le public comme une communauté agissante. Or, dans la plupart des projets technologiques, le public est réduit à la notion de consommateur, d'utilisateur ou d'usager lorsqu'il n'est pas purement et simplement disqualifié en tant que potentiel fraudeur, délinquant, ou terroriste (notamment dans les projets, nombreux, d'évaluation automatique des risques de fraude, de récidive, de radicalisation, de passage à l'acte…). Le « public » est une représentation projetée par les concepteurs, souvent sur la base de leurs propres « scripts » (Akrich 1992).

Or la construction éthique de valeurs communes ne relève pas une démarche marketing. En effet, une démarche éthique doit s'assurer de contribuer à la puissance d'agir des communautés concernées, c'est à dire garantir les modalités de leur participation aux décisions techniques qui les concernent (Zask 2008). Le marketing se contente le plus souvent de « sonder » des acteurs individuels pour tenter de les agréger comme une « masse » de consommateurs sans leur donner les moyens de s'organiser. A l'inverse, le rôle de la démarche éthique située est non seulement d'identifier des parties prenantes non prises en compte au départ, comme peuvent le suggérer les partisans de l'éthique externe (Lobet-Maris et al., 2019), mais aussi les constituer en public et de construire les modalités de leur participation effective à la sélection des projets, aux processus de conception comme à la gouvernance de l'usage des « solutions » qu'ils auront contribué choisir et à élaborer.

Cette approche se différentie sensiblement de l'éthique de la technologie à la Wright et de la plupart des initiative actuelles comme celle développée dans le rapport COMEST (Unesco,

2017). Elle est encore plus éloignée des raisonnements éthiques totalement abstraits et décorrélés des problèmes réels, comme peuvent l'être, par exemple, les nombreuses références au dilemme du tramway appliquées à la conduite autonome (par exemple, Tessier et al., 2018).

L'IA connexionniste et ses enjeux éthiques spécifiques

C'est l'IA connexionniste qui cristallise aujourd'hui l'essentiel des réflexions, des fantasmes solutionnistes, mais aussi des critiques (Morozov 2014). Les deux courants de l'IA, symbolique et connexionniste, ont aussi correspondus à deux grandes vagues de son développement, respectivement la deuxième vague des années 1980 à 2000 et la troisième, à partir des années 2010. D'un point de vue informatique, ces deux courants ont en commun, de recourir à des algorithmes utilisant des procédés dit heuristiques, c'est à dire efficaces dans un grand nombre de situations, mais toujours sujets à l'erreur parce qu'utilisant des « raccourcis ».

L'IA symbolique cherche à représenter de manière explicite les connaissances déclaratives de type « statiques » par des réseaux sémantiques ou des modèles objets et les connaissances procédurales par des règles d'inférence en utilisant des formalismes de type logique, même s'il ne s'agit pas forcément de logique formelle mathématique basée sur une sémantique vériconditionnelle, où le sens est ramené à la valeur de vérité des propositions. L'IA symbolique est utilisée dans presque tous les domaines de l'IA : système experts, planification, certains champs du traitement de la langue naturelle et certains domaines de l'apprentissage et surtout, plus récemment, en lien avec les sciences de l'information, dans le web sémantique et ses variantes (p.e le web socio-sémantique). L'IA symbolique implique une représentation explicite des objets et des activités qui permet de générer des « justifications » de la démarche suivie dans la résolution du problème.

L'IA connexionniste peut être assimilée à une forme de variante, assez profondément différente dans ses principes, de la statistique prédictive (Rouvroy 2013). Les algorithmes les plus connus sont ceux de l'apprentissage profond. Elle vise, en partant d'un ensemble de données, à les regrouper et à les classifier de manière ascendante sur la base d'un certain nombre de ressemblances. Mais à la différence de l'IA symbolique, les propriétés des éléments à comparer vont être progressivement enrichies, par essai-erreur, par des variables « cachées » qui ne correspondent pas à des attributs explicites d'entrée des « objets » comme pourraient l'être la taille, la couleur, la forme, l'âge, le genre, etc. Dans le cas des chaines de caractères, des images, des sons numérisés traités par les algorithmes d'apprentissage profond, les attributs

des « objets » sont représentés par des vecteurs de nombres. Au fur et à mesure de « l'apprentissage », la valeur de ces nombres est pondérée par d'autres vecteurs dans les couches cachées du réseau de neurone, sans qu'il soit possible ensuite de bien comprendre le poids des attributs initiaux des objets dans la décision. Paradoxalement, cette opacité peut rendre compte de l'aura de neutralité axiologique des modélisations algorithmiques (Rouvroy 2018).

Si les principes algorithmiques de l'IA connexionniste sont très anciens, elle a connu un renouveau considérable ces dernières années grâce à la mise à disposition de données massives (les big data) issues de la traçabilité de très nombreuses activités via les applications du web, des smartphones ou par l'exploitation des grandes bases de données de gestion des entreprises et des administrations. De fait, les applications de l'IA connexionniste ont effectivement des effets dans la vie quotidienne de nombreux consommateurs et citoyens du fait de la tendance à l'automatisation de très nombreuses interactions de service. Citons, par exemple, des applications qui recourent à ces divers procédés de la statistique prédictive, très souvent basées aujourd'hui sur l'IA connexionniste :

- la notation et les scores de risque, par exemple, attribution de scores de risques de récidive à des candidats à la libération conditionnelle ou à des prévenus en attente de procès (algorithme COMPAS) ; scores de risques de non remboursement de prêts bancaires ; scores de risques de fraude ; social credit scoring chinois ; système de notation des travailleurs des plateformes par les utilisateurs ; systèmes d'attribution de scores de risque d'être impliqué à titre d'auteur ou de victime d'actes de violence (algorithme PREDPOL)… ;
- les appariements (exemples : bob emploi ; parcoursup ; sites de rencontre ;…) ;
- La hiérarchisation (PageRank de Google, EdgeRank de Facebook…) ;
- La personnalisation des offres commerciales (Amazon ; Target…) ou de contenus médiatiques (Netflix…) ; le ciblage du marketing politique (campagnes électorales, propagande) fondé sur le profilage psychographique (Cambridge Analytica) ;
- La géolocalisation et la fluidification des déplacements dans les espaces publics (Waze, Google Maps,…)
- La traduction automatique (Google translate n'est plus fondé sur un raisonnement « réductionniste » formel combinant règles syntaxiques abstraites et dictionnaires, mais sur la détection statistique et contextuelle « évolutionniste » de corrélations dans les corpus numérisés)

- La domotique (internet des objets).

Quand bien même chaque type d'algorithme ou d'application présente des enjeux spécifiques, ils ont tous en commun le fait d'être « conduits par des données » (data-drivenness), c'est à dire d'être fondées sur le traitement automatisé de données numériques plutôt que sur des règles conventionnelles explicites, ou des normes ou consensus politiquement ou collégialement débattus et contestables dans une forme d'opérationnalité sans « épreuve[2] ». Le recours à des algorithmes de prédiction opaques et exploitant des sources de données difficilement interprétables d'un point de vue social mais pouvant néanmoins contribuer à la décision automatique, contribue à renforcer l'impression que ces données « brutes », à l'inverse des données signifiantes, correspondent à de purs signaux, à un « langage des choses » émanant « spontanément » du monde, qu'elles ne sont pas « produites » et reflètent donc objectivement le monde en soi.

C'est sur ces prétentions d'objectivité et d'impartialité algorithmique que se focalisent les « critical data studies » (Iliadis & Russo 2016) :

« *les Critical Data Studies (CDS) explorent les défis culturels, éthiques et critiques uniques que posent les Big Data. Plutôt que de traiter les Big Data comme des phénomènes uniquement empiriques sur le plan scientifique et donc largement neutres, les CDS défendent l'idée que les Big Data doivent être considérés comme des ensembles de données toujours constitués au sein d'ensembles de données plus larges. Le concept d'assemblages permet de saisir la multitude de façons dont les structures de données déjà constituées infléchissent et interagissent avec la société, son organisation et son fonctionnement, et l'impact qui en résulte sur la vie quotidienne des individus. Le CDS remet en question les nombreuses hypothèses sur les Big Data qui imprègnent la littérature contemporaine sur l'information et la société en repérant les cas où les Big Data peuvent être naïvement considérées comme des entités informationnelles objectives et transparentes* ».

En effet, de nombreuses discriminations algorithmiques bien recensées dans la littérature sont commises : algorithme de recrutement reproduisant les biais favorables aux hommes, mauvaise reconnaissance des femmes de couleur par les algorithmes de reconnaissance faciale, facteur de risque dans l'étude de maladies moins précis pour les patients d'origine africaine ou asiatique, algorithme de détection des risques d'implication, à titre de victime ou d'acteur, dans des faits de violence, ou algorithme d'évaluation et de notation des risques de récidive générant des résultats faussement positifs plus souvent pour les noirs américains que pour les Blancs,

etc. Ceci est d'autant plus problématique que le fonctionnement des algorithmes que nous avons évoqué plus haut, qui tend à créer des variables intermédiaires cachées (« proxies ») difficilement repérables et interprétables, rend à la fois la décision difficile à justifier et son caractère éventuellement discriminatoire difficile à prouver.

Mais au-delà – ou en deçà - des prétentions d'objectivité et d'impartialité propres à l'idéologie technique de la « data-drivenness », la signification politique de la « fonction objective » - c'est-à-dire de la logique sectorielle que les concepteurs ont choisi de maximiser en la formalisant sous la forme de « contraintes » devant borner les opérations d'optimisation – est assez systématiquement oblitérée sous des arguments de « rationalisation », d'accélération et de facilitation d'accès. Ainsi, le Rapporteur spécial des Nations Unies sur l'extrême pauvreté et les droits humains (Alston 2019), documentait-il la manière suivant laquelle - aux Etats-Unis et en Angleterre notamment - l'introduction de la décision algorithmique dans la gestion de l'aide sociale (automatisation du traitement des demandes, de la détection des fraudes ou risques de fraudes…) a eu pour conséquence d'accentuer la précarité des couches les plus vulnérables de la population.

C'est que l'État providence algorithmique est « gouverné » par des fonctions objectives aisément modélisables de rationnement et d'austérité plutôt que de mise en œuvre des principes beaucoup moins aisément « calculables » d'indivisibilité, d'interdépendance et d'indissociabilité des droits économiques et sociaux et des droits civils et politiques (principes constamment réaffirmés par les Nations Unies et le Conseil de l'Europe). Dans le cas de l'État providence algorithmique dénoncé par le Rapporteur spécial des Nations Unies, la conjonction des algorithmes et d'une bureaucratie rendue amnésique des fondements de son action[5] transforme *de facto* les droits économiques et sociaux en variables d'ajustement conditionnées à l'atteinte d'objectifs toujours plus éloignés de croissance économique.

Ce ne sont pas que les « données » qui ne sont pas « données » : la détermination des « finalités » et des « contraintes » de l'optimisation ne peuvent pas non plus être sous-traitées aux seuls concepteurs techniques des algorithmes : en tant qu'ils concernent les critères de

[5]Notons toutefois que, dans son rapport, Philippe Alston fait également état de la situation en Ontario, où, en raison d'erreurs massives du système algorithmique d'évaluation de l'éligibilité des personnes demandeuses d'aide sociale, des fonctionnaires en charge usèrent d'une série de subterfuges afin de garantir un traitement équitable des demandes.

mérite, de besoin, de désirabilité, de dangerosité présidant à la répartition des opportunités et des ressources, ils relèvent fondamentalement d'une théorisation de la justice qui ne peut être déterminée que collectivement, suivant les formes prescrites en démocratie délibérative.

Les approches externalistes de l'éthique, correspondent assez bien aux propositions formulées dans l'article de Besse et collaborateurs (Besse et al. 2019), qui considèrent que ce défaut des algorithmes connexionnistes doit être traité, d'une part sur le plan des valeurs et d'autre part d'un point de vue technique, en apportant divers correctifs. Sur le plan des valeurs, il s'agit de renforcer l'arsenal juridique ou de s'assurer de sa bonne mise en œuvre. Sur le plan technique, il s'agit d'améliorer la qualité des données ou de travailler à améliorer l'explicabilité. Or, comme nous le verrons en adoptant une approche d'éthique située qui déploie d'autres perspectives disciplinaire[6], les problèmes de qualité des données comme d'explicabilité sont inhérents au fonctionnement de l'IA connexionniste et ne peuvent pas faire l'objet de correctifs qui remettraient fondamentalement en cause ces principes. Sur le plan des valeurs, qui renvoie à la subjectivité du public et aux normes qu'il soutient ou qu'il remet en cause en lien avec la construction des faits nouveaux mis en lumière par la pluridisciplinarité, cela signifie qu'il est tout simplement impossible de sous-traiter des décisions à fort impact humain et social à des dispositifs automatiques utilisant des heuristiques, aussi apprenants ou autonomes soient-ils, quel que soit par ailleurs l'arsenal juridique déployé pour les « encadrer ».

Conclusion

Le recours à l'éthique dans les applications de l'IA est souvent justifié par les tenant d'une IA forte pour atténuer les effets de son « intelligence » considérée comme un acquis scientifique entrainant le caractère inéluctable de son développement. Les approches classiques de l'éthique de la technologie acceptent ce rôle en tentant d'atténuer l'impact de la « technologie inéluctable » sur les bénéficiaires supposés.

Notre vision d'une éthique située, basée sur le pragmatisme et la philosophie des sciences, assigne à l'éthique un rôle très différent. Celui-ci consiste à remettre en cause les présupposés de scientificité et de neutralité axiologique qui justifient le recours à la technologie en ouvrant des espaces de controverses scientifiques et politiques masquées, telles des boites noires, par

[6] Cf. notamment notre texte en préparation pour la RFSIC.

les promoteurs de la technologie inéluctable. L'intervention de l'éthique a alors toujours pour conséquence de redéfinir les contours du projet, de ses finalités comme de ses avantages attendus, dans une veine de design critique qui peut notamment contribuer à des éléments de prospectives[7] ou d'ouverture vers d'autres possibles.

Car nous ne pensons pas que le parti pris de l'éthique située soit celui d'un refus radical et systématique de la technologie. Nous pensons, conformément à une certaine épistémologie des SIC, qu'il s'agit toujours de montrer comment les technologies de l'information et de la communication, dans la prolongation de l'écriture, sont des télé-technologies pour reprendre l'expression de Derrida (Delain 2006), ou encore des pharmakons, à la fois remède et poison, toujours selon Derrida à suite de Platon, dans une veine actualisée par Stiegler (2007), dont il faut concevoir les usages avec prudence. Cette prudence invite à suivre la voie d'une forme de démocratie technique (Callon et al. 2014) ou pour le dire dans les termes de Stengers de réactiver le « sens commun » (Stengers 2020) dans les projets d'innovation, en renvoyant dos à dos les assertions solutionnistes des promoteurs de la « technologie inéluctable » et celles des partisans du refus obstiné du changement technique considéré comme nécessairement déshumanisant.

Remerciement

Nous remercions Etienne-Armand Amato pour sa relecture attentive du manuscrit.

Bibliographie

Alston, P. (2019). Report of the Special Rapporteur on extreme poverty and human rights, submitted in accordance with Human Rights Council resolution 35/19, Seventy-fourth session, A/74/48037, 11 octobre 2019.

Akrich, M. (1992). The De-scription of Technical Objects. Dans W. E. Bijker et j. Law (dir.), *Shaping technology/Building Society. Studies in Sociotechnical Changes* (p. 205-224). Cambridge: MIT Press.

Baron, X. et Cugier, N. (2016). Des « Services généraux » aux « aménités » des environnements du travail. *L'expansion Management Review*. http://www.bmvr.nice.fr/EXPLOITATION/Default/doc/ALOES/4875556/services-generaux-aux-amenites-des-environnements-du-travail-des

[7] Au sens de Gaston Berger (Berger et al. 2007), la prospective, au lieu de consister – comme pour les « prédicateurs » de la Silicon Valley – sur l'extrapolation au départ de tendances du passé, comme la soi-disant loi de Moore qui n'a rien d'une loi, consiste à imaginer, au contraire, des ruptures relativement à l'état de fait, en fonction d'un horizon de perfectibilité du social qui ressemble fort à l'idée de la justice.

Beauchamp, T. L. et Childress, J. F. (2019). *Principles of biomedical ethics* (Eighth edition). Oxford University Press.

Berger, G., Bourbon Busset, J. de et Massé, P. (2007). *De la prospective textes fondamentaux de la prospective française 1955-1966* (Deuxième édition; édité par P. Durance). L'Harmattan.

Besse, P., Castets-Renard, C., Garivier, A. et Loubes, J.-M. (2018, octobre). *L'IA du quotidien peut-elle être éthique ?* https://hal.archives-ouvertes.fr/hal-01886699

Bonnemains, V., Tessier, C. et Saurel, C. (2018). Machines autonomes « éthiques » : questions techniques et éthiques. *Revue francaise d'ethique appliquee*, *N° 5*(1), 34-46. http://www.cairn.info/revue-francaise-d-ethique-appliquee-2018-1-page-34.htm

Bordage, F. (s. d.). *Empreinte environnementale du numérique mondial*. https://www.greenit.fr/empreinte-environnementale-du-numerique-mondial/

Callon, M., Lascoumes, P. et Barthe, Y. (2014). *Agir dans un monde incertain essai sur la démocratie technique* (Édition révisée). #0, Éditions Points.

Calude, C.S., Longo, G. (2017). "The Deluge of Spurious Correlations in Big Data." Found Sci 22, 595–612. https://doi.org/10.1007/s10699-016-9489-4

Clot, Y. et Stimec, A. (2013). « Le dialogue a une vertu mutative », les apports de la clinique de l'activité. *Negociations*, *n° 19*(1), 113-125. http://www.cairn.info/revue-negociations-2013-1-page-113.htm

Conseil d'Orientation pour l'Emploi. (2017). *Automatisation, numérisation et emploi - Tome 1*. https://www.strategie.gouv.fr/sites/strategie.gouv.fr/files/atoms/files/coe-rapport-tome-1-automatisation-numerisation-emploi-janvier-2017.pdf

Conseil d'État, L. C. (s. d.). *Transparence, valeurs de l'action publique et intérêt général*. Conseil d'État. https://www.conseil-etat.fr/actualites/discours-et-interventions/transparence-valeurs-de-l-action-publique-et-interet-general

Delain, P. (2006). *Derrida, technosciences, télé-techniques, médias*. https://www.idixa.net/Pixa/pagixa-0611051647.html

Dewey, J. (1927). *Le public et ses problèmes* (traduit par J. Zask). Gallimard.

Dewey, J. (2008). La théorie de la valuation. *Tracés. Revue de Sciences humaines*, (15), 217-228. 10.4000/traces.833

Domenget, J.-C. et Wilhelm, C. (2017). Un nécessaire questionnement éthique sur la recherche à l'ère des Digital Studies. *Revue française des sciences de l'information et de la communication*, (10). 10.4000/rfsic.2668

Eubanks, V., *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor,* St-Martin's Press, 2018.

Flipot, F., Dobré, M. et Michot, M. (2013). *Livre : la face cachée du numérique - Green IT*. L'échappée. https://www.greenit.fr/2019/05/22/livre-la-face-cachee-du-numerique/

Frey, C. B. et Osborne, M. A. (2013). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254-280. 10.1016/j.techfore.2016.08.019

Ganascia, J.-G. (2019). Le mythe de la Singularité faut-il craindre l'intelligence artificielle ? #0, Éditions Points.

Guattari, F. (1989). *Les trois écologies*, Galilée. https://static1.squarespace.com/static/5657eb54e4b022a250fc2de4/t/566fa0cddf40f39ea7f3d8bb/1450156237851/1989_F%C3%A9lix+Guattari_Les+Trois+Ecologies.pdf

Hache, É. (2011). *Ce à quoi nous tenons*. La Découverte. 10.3917/dec.hache.2011.01

Hopkins, R., Ponticelli, A. et Vermeersch, L. (2017). Everything gardens : les villes en transition. *Vacarme*, *N° 81*(4), 28-38. https://www-cairn-info.proxybib-pp.cnam.fr/revue-vacarme-2017-4-page-28.htm

Iliadis, A. et Russo, F. (2016). Critical data studies: An introduction. *Big Data & Society*, *3*(2), 2053951716674238. 10.1177/2053951716674238

Lobet-Maris, C., Grandjean, N., Vos, N. D., Thiry, F., Pagacz, P. et Pieczynski, S. (2019). Au cœur de la contrainte : quand l'éthique se fait bricolage. *Revue francaise d'ethique appliquee*, *N° 7*(1), 72-88. https://www-cairn-info.proxybib-pp.cnam.fr/revue-francaise-d-ethique-appliquee-2019-1-page-72.htm

Morozov, E. (2014). *Pour tout résoudre, cliquez ici : l'aberration du solutionnisme technologique*. Fyp éditions. https://bibliotheques.paris.fr/Default/doc/SYRACUSE/980590/pour-tout-resoudre-cliquez-ici-l-aberration-du-solutionnisme-technologique

Muniesa, F. et Callon, M. (2013). 8. La performativité des sciences économiques. Dans P. Steiner et F. Vatin (dir.), *Traité de sociologie économique* (p. 281-316). Presses Universitaires de France. http://www.cairn.info/traite-de-sociologie-economique--9782130608318-page-281.htm

Ochigame, R. (2019). The Invention of "Ethical AI": How Big Tech Manipulates Academia to Avoid Regulation. *The Intercept*. https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/

Parisi, L. (2017). « Reprogramming Decisionism », e-flux journal, n.85.

Parisi, L. (2019). The alien subject of AI. *Subjectivity*, *12*(1), 27-48. 10.1057/s41286-018-00064-3

Rouvroy, A. (2013). "The End(s) of Critique: Data-behaviourism vs. Due Process", in. M. Hildebrandt and K. De Vries (eds.), Privacy, Due Process and the Computational Turn, London: Routledge, 143-168.

Rouvroy, A. (2016), Des données et des Hommes. Droits et libertés fondamentaux dans un monde de données masssives. Conseil de l'Europe, T-PD-BUR (2015) 09REV.

Rouvroy, A. (2017), "Contestability in the Big Data Era", Archival Ucertainty. Workshop on #Error, Department of Arts and Cultural Studies, University of Copenhagen, Københavns Stadsarkiv, Copenhagen City Hall, 14 November, 2016. https://ku.23video.com/search/perform?search=uncertain+archives

Rouvroy, A. (2018a), "homo juridicus est-il soluble dans les données ?", Droit, Normes et Libertés dans le Cybermonde, Larcier, 417-444.

Rouvroy, A. (2018b), "Mapping as governance in an Age of Autonomic Computing: Technology, Virtuality and Utopia" in P.Barguès-Pedreny, D. Chandler, E. Simon (dir.), Mapping and Politics in the Digital Age, London: Routledge, p.118-134.

Rouvroy, A. (2020), "L'usage des "Big Data" pour gouverner", Politique (numéro special Covid19:tout repenser. La pandémie, miroir des inégalités'), no.112, 115-119.

Serikoff, G., Foliot, C. et Zacklad, M. (2019). *Le Lab des Labs*. https://hal.archives-ouvertes.fr/hal-02437318

Stengers, I. (2020). Réactiver le sens commun lecture de Whitehead en temps de débâcle. #0, Éditions La Découverte.

Stiegler, B. (2007). Questions de pharmacologie générale. Il n'y a pas de simple pharmakon. *Psychotropes*, *Vol. 13*(3), 27-54. http://www.cairn.info/revue-psychotropes-2007-3-page-27.htm

Unesco, Commission d'éthique des cns. scientifique et technos. (2017). *Report of COMEST on robotics ethics - UNESCO Bibliothèque Numérique*. https://unesdoc.unesco.org/ark:/48223/pf0000253952

Villani, C. (2017). *Donner un sens à l'Intelligence Artificielle*. https://www.aiforhumanity.fr/pdfs/MissionVillani_Presse_FR-VF.pdf

Wright, D. (2011). A framework for the ethical impact assessment of information technology. *Ethics and Information Technology*, *13*(3), 199-226. 10.1007/s10676-010-9242-6

Zacklad, M. (2012). Vers une informatique au service de l'homme. *Personnel. La revue de l'ANDRH*, 63-64. https://halshs.archives-ouvertes.fr/halshs-02937484

Zacklad, M. (2018). *Intelligence Artificielle : représentations et impacts sociétaux* ([Raport technique]). CNAM. https://halshs.archives-ouvertes.fr/halshs-02937255

Zacklad, M. (2019). Entretien (Design, conception, création, Vers une théorie interdisciplinaire du design). *MEI : Information et Médiation*, (41), 9-31. https://mei-info.com/revue/41/6/entretien/

Zacklad, M. (2020). Les enjeux de la transition numérique et de l'innovation collaborative dans les mutations du travail et du management dans le secteur public. Dans *A. Gillet (Éd.), Les transformations du travail dans les services publics (Presses de l'EHESP)*. https://hal.archives-ouvertes.fr/hal-02934479

Zask, J. (2008). Le public chez Dewey : une union sociale plurielle. *Tracés. Revue de Sciences humaines*, (15), 169-189. 10.4000/traces.753