L’éthique située de l'IA et ses controverses

Rouvroy, Antoinette; Zacklad, Manuel

Published in:
Revue Française des sciences de l'Information et de la Communication

Publication date:
2022

Document Version
Version revue par les pairs

L’éthique située de l'IA et ses controverses

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

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

Mots-clefs : Éthique, Éthique située, Intelligence Artificielle, Apprentissage profond, Controverse

Keywords: Ethic, Situated ethic, Artificial Intelligence, Deep Learning, 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. Enfin, nous introduirons les six espaces de controverses de l’éthique située de l’IA.

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. Finally, we will introduce the six space of shared controversy of the situated ethics of connectionist AI.

L’éthique située de l'IA et ses controverses

Manuel Zacklad
Antoinette Rouvroy

Introduction

Dans cette communication[1][2], 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 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 2018) 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[3]. Enfin, nous introduirons les six espaces de controverses de l’éthique située de l’IA[4].

---

[1] Cet article 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.

[2] Cet article correspond également à la version longue de la présentation au congrès de la SFSIC 2021 intitulée « Enjeux éthiques situés de l'IA »

[3] 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)

[4] Qui pour certaines concernent aussi plus largement une éthique située du numérique.

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 (2019), 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…), de la 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, la protection de la vie privée et les principes de protection des données à caractère personnel tels qu’ils sont, dans le contexte européen, inscrits dans la Convention européenne des droits de l’Homme, dans la Charte des droits fondamentaux de l’Union européenne et dans le RGPD notamment.

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 (warranted assertibility) par des expériences répétées (Dewey 1941). 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. généralement 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, la production ou le renforcement, à travers la "décision algorithmique", d'inégalités de moyens, d'opportunités ou de considération affectant défavorablement des portions de populations juridiquement protégées contre la discrimination fondée sur le genre ou sur l'origine ethnique (notamment), est à l'évidence un problème juridique. Mais comment apporter la preuve de discriminations indirectes, c'est-à-dire de distinctions de traitement qui, sans être explicitement fondées sur des critères de distinction interdits par le droit, ont néanmoins pour effet de défavoriser significativement les membres d'un groupe

« minoritaire » ? Ces décisions sont cependant considérées comme soulevant un problème éthique parce qu'il y a une dissonance entre la sorte de « réalisme naïf » imprégnant la croyance en l'adéquation et l'exhaustivité, la neutralité et l'impartialité des données et l'objectivité machinique des algorithmes qui les exploitent pour fournir des décisions automatiques, d'un côté, et les effets discriminatoires constatés, de l'autre, et interprétés comme des « erreurs » au regard de la croyance en l'objectivité machinique. Présenter les « biais » algorithmiques comme un problème susceptible de trouver sa solution dans une approche « ethics by design », n'est en fait rien d'autre qu'une concession stratégique qui ne met aucunement en question la croyance en la possibilité « d'algorithmes éthiques » qui auraient pour « propriétés » d'être véritablement objectifs, impartiaux, équitables, justes…

Gilles Deleuze (Deleuze 1968 : 193) avait très exactement exprimé que « [L]'erreur n'est que l'envers d'une orthodoxie rationnelle, et elle témoigne encore en faveur de ce dont elle s'écarte, en faveur d'une droiture, d'une bonne nature et d'une bonne volonté de celui qui est dit se tromper ». Parler d'erreur présuppose une certaine idée de la vérité, dont on a bien compris qu'elle n'est pas accessible et qu'elle n'est pas l'enjeu. Donc plutôt que de parler d'erreur, qui renvoie inévitablement et renforce l'idée de l'existence et de l'accessibilité d'une « vérité vraie », d'une objectivité parfaite, qui nierait le fait « qu'il n'est pas possible d'isoler un moment de la mesure qui serait indépendant de ses usages » (Desrosières 2014), nous parlerons plutôt de « contestabilité ». C'est la possibilité de les contester qui assure pragmatiquement une robustesse et une légitimité suffisante aux productions statistiques, aux modélisations, qui n'ont aucune vocation à être « vraies » ni « fausses », l'enjeu se situant plutôt du côté du « crédit » qu'elles « produisent », ou de la « fiabilité » qui dispense, précisément, des opérations de véridictions qui seraient de rigueur relativement à des énoncés autres que statistiques, ou algorithmiques (des énoncés fondés sur l'observation directe, empirique, des phénomènes physiques par exemple).

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. En effet, la transposition, dans « l'univers » des données massives et des algorithmes

d'apprentissage machine, du vocabulaire convenu de la statistique est tout sauf évidente,[5] puisque, dans ce « nouveau » contexte, ce qui tient lieu de « mesure », c'est la collecte ou l'enregistrement - très peu sélectif - des données qui ne relève pas de conventions de quantification établies ni de distinction *a priori* entre les « signaux » et les « bruits ». Quant aux usages que, dans le droit de la protection des données, on appelle les « finalités », dont Alain Desrosières affirmait qu'ils conditionnent nécessairement la mesure, ils ne sont tout simplement pas déterminés avant la collecte, mais plutôt dans « l'après coup » d'une temporalité marquée par la récursivité. Voilà pourquoi les dispositifs algorithmiques apprenants nourris aux données massives mais surtout rapides (*fast data*) sont plutôt voués à la dé-mesure (qui convient parfaitement à la dérégulation), la dés-intelligence des limites propres au capitalisme néolibéral dans ses dernières recombinaisons en date, et à l'in-gouvernance (accélération de la fragmentation, de la dérégulation, de l'entropie) (Rouvroy 2016b).[6]

Cette remise en cause des présupposés scientifiques des promoteurs de la décision automatisée ne peut pas être le fait d'éthiciens professionnels qui seraient les garants de valeurs transcendantes mais résulte de la capacité à poser autrement les problèmes de nature, sémiotique, épistémique, 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 d'optimisation « 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

---

[5] Comme nous le faisait remarquer l'un des reviewers de la présente contribution.

[6] Nick Land affirmait, dans le même ordre d'idée, depuis la perspective nihiliste d'un « accélérationnisme de droite », l'identité téléologique entre le capitalisme et l'intelligence artificielle (Land, 2014).

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 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 d'audience, 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 d'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 sont aussi toujours des décisions politiques – 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 de 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 à choisir et à élaborer.

Cette approche se différencie 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 abstraits et dé-corrélés des problèmes réels, comme peuvent l'être, par exemple, les références au dilemme du tramway dans le contexte de la conduite autonome (par exemple, Bonnemains et al., 2018).

L'IA connexionniste et ses enjeux éthiques spécifiques

Avant d'aborder les six espaces de controverses qui caractérisent l'éthique située de l'IA et du numérique, revenons un moment sur la spécificité du moment actuel de l'IA connexionniste qui cristallise l'essentiel des réflexions, des fantasmes solutionnistes (Morozov 2014), mais aussi des critiques. Ces 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 fréquentiste (Rouvroy 2013). Plutôt que de viser à rendre lisibles des « régularités » (comme des courbes de Gauss), l'IA connexionniste, met en œuvre,

pour inférer des probabilités, non plus aucune hypothèse fréquentiste (aucune catégorie statistique conventionnellement pré-déterminée), mais la détection de corrélations dans l'ensemble de tous les points de données disponibles pour produire des inférences sur la probabilité d'occurrences futures basées sur des exemples passés, et l'optimisation des résultats initiaux sur la base des observations ultérieures. Contrairement aux statistiques fréquentistes, les probabilités bayésiennes ne cherchent pas à rendre sensibles des régularités, mais plutôt à créer des processus d'optimisation. Elle vise, en partant d'un ensemble de données, à les regrouper et à les classifier de manière ascendante sur la base de détections non pas de ressemblances mais de proximités, non pas de de rapports de causes à effets mais de corrélations. Dans le cas des chaines de caractères, des images, des sons numérisés traités par les algorithmes d'apprentissage profond, les « attributs » (qui ne sont plus véritablement des attributs mais plutôt des « signaux ») des « objets » sont représentés par des vecteurs de nombres. Au fur et à mesure de « l'apprentissage », le poids relatif de ces signaux est pondéré par d'autres vecteurs dans les couches cachées du réseau de neurones, sans qu'il soit possible ensuite de bien comprendre le poids des signaux initiaux des objets dans la décision. Ainsi, l'importance relative des données personnelles ou biographiques d'une personne affectée par un profilage, un appariement, la détermination d'un score de risques ou de crédit peut être minime, voir même nulle. Grâce aux algorithmes fonctionnant sur une logique d'induction bayésienne, il est possible de caractériser relativement finement les comportements possibles d'une personne sans traiter aucune donnée à caractère personnel. Paradoxalement, cette opacité peut contribuer à l'aura de neutralité axiologique des modélisations algorithmiques (Rouvroy 2018a).

Si les principes algorithmiques de l'IA connexionniste sont très anciens, elle a connu un renouveau considérable ces dernières années grâce à la disponibilité de fait 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 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 fondés 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 ». 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 des données massives 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, la transcription ou l'enregistrement de la réalité sociale sous forme de données numériques ne la purge bien évidemment pas des inégalités mais les « naturalise », faisant passer les données pour des « faits » en faisant oublier que les « faits » sont toujours produits, et que les données ne traduisent jamais que les « effets » des rapports de force et des phénomènes de domination. L'amnésie des conditions de production des données, de leur contexte référentiel, de leur conditionnalité a été maintes fois dénoncée, et la littérature a objectivé la responsabilité de ce déni dans la production et la perpétuation de stigmatisations et discriminations : 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, ne mobilise explicitement et directement aucun « critère » illégal de discrimination repérable comme tel. Par exemple, plutôt que l'origine ethnique, ce sera bien plutôt des données révélant le code postal du domicile, corrélé à un « risque » - un « risque », ce n'est encore « personne » - accru d'être impliqué, en tant qu'auteur ou victime, dans des actes de violence, qui interviendra en « remplacement » des données ethniques, alors que les codes postaux peuvent indirectement révéler, avec une haute probabilité, l'origine ethnique des personnes qui y habitent, sans mobiliser aucune donnée de « genre ». De même, l'algorithme de recommandation à l'embauche entraîné sur base des données fournies par les employeurs dans les secteurs dans lesquels persiste le « plafond de verre » empêchant les femmes d'accéder aux postes les mieux rémunérés, répercutera et perpétuera passivement, à travers ses recommandations, cet état de fait. La « discrimination indirecte » ainsi systématiquement produite est de fait difficilement repérable et interprétable, alors que le caractère de « boîte

noire » des algorithmes apprenants rend aussi bien la « justification » de la décision que sa « contestation », extrêmement difficiles.

Les approches externalistes de l'éthique, correspondent assez bien aux propositions formulées dans l'article de Besse et collaborateurs (Besse et al. 2018), 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, 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 ».

Six espaces de controverses de l'éthique du numérique et de l'IA

Nous décrivons six domaines de l'éthique situées faisant l'objet de débats scientifiques et de conflit de valeurs. Chacun remet en cause des assertions scientifiques en ouvrant un espace de controverses sur l'IA et ses usages, qui au-delà des domaines de l'informatique et du droit convoque un grand nombre de sciences humaines et sociales allant des SIC à la philosophie des sciences en passant par la sociologie, l'ergonomie et la psychologie du travail. Ces espaces de controverses concernent souvent les nouveaux développements associés à l'essor de l'IA connexionniste mais ils sont généralement valables pour l'ensemble de l'IA d'autant qu'IA connexionniste et IA symbolique sont associées dans plusieurs applications[7].

A. Controverses liées aux enjeux de culture numérique et de croyances liées à l'autonomie des IA et à l'objectivité des données

[7]Comme le souligne un de nos reviewer.

Le débat éthique a d'abord et avant tout sa source dans les imaginaires et les croyances liées aux technologies qui sont en partie dépendants des travaux scientifiques dans le domaine de l'informatique et du management notamment. Dans le domaine de l'IA, le débat se cristallise sur l'opposition entre l'IA faible et l'IA forte. Celle-ci est notamment basée sur le postulat selon lequel les recherches en informatique doteront bientôt les machines d'une authentique conscience et d'une intelligence supérieure qui pourrait dépasser celle de l'homme, ce qui justifie dès à présent leur utilisation tout le temps et partout pour la performance des organisations et le bien-être de l'humanité. L'IA forte s'accompagne de certaines craintes pour le moins farfelues selon lesquelles les IA, supérieures à l'homme en intelligence, pourraient prendre leur autonomie pour le concurrencer voire le menacer (Ganascia 2019). Le courant est aussi alimenté par les mythes souvent délirants du transhumanisme, comme celui du transfert de l'esprit d'une personne dans un support artificiel, mythes relayés dans la presse grand public avec un verni d'autorité scientifique fournit par des universités privées financées par certaines grandes entreprises du numérique, telle l'Université de la Singularité.

Cette vision est elle-même appuyée de manière plus subtile sur deux autres positions. La première est celle d'une vision substantialiste de l'intelligence voire de la créativité. Ces dernières seraient des attributs intrinsèques des individus basées sur les caractéristiques de leur cerveau et il serait donc totalement possible de reproduire ces propriétés dans des programmes. Les tentatives, assez pathétiques, de rechercher dans l'anatomie du cerveau d'Einstein la source de sa créativité, indépendamment du milieu intellectuel dans lequel il baignait, relèvent de cette vision.

La deuxième, qui concerne l'IA connexionniste et l'apprentissage profond, est basée sur le postulat selon lequel l'accès à des données neutres et de grande qualité permettrait d'obtenir des performances « objectives » supérieures à celle de l'homme ce qui, en particulier, règlerait les problèmes de « biais » considérés comme un défaut de jeunesse de ces algorithmes. Derrière cette représentation réside l'idée selon laquelle l'objectivité calculatoire est supérieure à la subjectivité des acteurs humains soumis eux aussi à des biais[8] cognitifs et émotionnels quasi insurmontables.

[8] Selon une autre perspective on pourrait considérer que sans subjectivité assumée, donc sans biais, il n'y a pas d'apprentissage possible.

Sur ce point, les éléments de la controverse portent sur la croyance dans les données objectives. Pour les tenants de la thèse opposée, dans la réalité sociale, comme le disait Latour (2007), il ne faut ne pas considérer les données comme « données » mais comme « obtenues ». Cela signifie, premièrement, que les « obtenues » sont tributaires de la qualité (sensibilité) et de la distribution spatiale des « capteurs » et des « émetteurs » de données numériques. Le cas des applications de traçage des contacts à des fins sanitaires via des signaux bluetooth est exemplaire à cet égard : les téléphones les plus sophistiqués émettent en général un signal plus « fort » que les téléphones bon marché…dont les signaux bluetooth seront moins bien détectés (Rouvroy 2020).

Deuxièmement, cela signifie aussi, comme déjà évoqué plus haut, que les « obtenues » ne reflètent pas tant les « faits » que les « effets » des rapports de force et de domination dont la réalité sociale est parcourue. Par exemple, les bases de données d'entraînement de l'algorithme PREDPOL, qui contiennent des données relatives à toutes les personnes ayant fait l'objet d'une interpellation policière – fût-ce pour un simple contrôle d'identité – sont caractérisées par une surreprésentation de la population Noire américaine…l'une des raisons étant que les afro-américains sont plus souvent contrôlés que les caucasiens et plus souvent incarcérés aussi…cette surreprésentation dans les données d'entraînement « naturalise » ce qui, pour une part au moins, est le résultat de stigmatisations et de discriminations historiques. Le même phénomène existe pour les algorithmes de recrutement qui favorisent les hommes en reproduisant leur surreprésentation à des postes d'encadrement. Les données, même qualitatives, enregistrent fidèlement l'état de fait résultant de ces pratiques, que les algorithmes reproduisent et « naturalisent », les rendant moins aisément décelables et contestables (Rouvroy 2018a).

De fait, les modèles dits prédictifs sont toujours affectés d'un « compromis » entre l'adéquation du « modèle » aux données d'entraînement (et donc à la reproduction des « biais » transcrits sous forme de données et à la modélisation performative-consrvatrice du monde social) et ce que l'on appelle la « variance » (la propension du « modèle » à changer dès-lors que varient les données en entrée). Les modèles les plus simples ont une faible variance mais conservateurs des « biais » contenus dans les données d'apprentissage. Les modèles plus complexes sont d'avantage prédisposés à ce que l'on appelle les phénomènes d' «over-fitting » ou de « sur-apprentissage ». Le modèle « apprend » à tellement bien analyser les données d'entrainement qu'il prend pour des « concepts » des fluctuations arbitraires dans ces données, ce qui rend le modèle très peu capable de se « généraliser », c'est-à-dire de produire des résultats fiables dès

lors qu'il serait exposé à de nouvelles données, différentes de celles qui ont servi à son apprentissage (Geman et al. 1992). C'est le problème des « spurrious correlations » (Calude & Longo 2017), la détection de patterns qui ne sont dans les données massives que par le pur effet du hasard. Ce problème des corrélations abusives est d'ailleurs le propre de l'interprétation de tous les résultats produits par des méthodes inductives. Par exemple, lorsque A et B sont fortement corrélés, cela peut signifier que « A cause B », que « B cause A », mais aussi que C, que l'on n'a pas détecté, « cause à la fois A et B ». Ce « compromis » explique aussi « l'arbitrage[9] » existant entre explicabilité et fiabilité : un modèle qui varie peu et reste « fidèle » aux données d'entrainement peut être plus facilement explicable qu'un modèle très dynamique, qui varie en fonction des données qu'il rencontre, mais un modèle très « stable » n'est pas nécessairement adéquat pour rendre compte de ce qui s'en écarte, alors qu'un modèle très « dynamique » peut apparaître plus « fiable » de ce point-de-vue.

Du coté des traitements automatisés, les biais résident aussi, bien évidemment, dans la « fonction objective » des algorithmes d'optimisation qui détermine « ce que » l'algorithme est censé optimiser :il peut s'agir par exemple, sur un site de librairie en ligne, de favoriser les best-sellers au détriment des ouvrages plus singuliers ou encore, dans les applications de prévention du terrorisme, de tolérer un nombre important de victimes collatérales et de faux positifs pourvu que le passage à l'acte terroriste soit évité (Rouvroy 2018b). Les algorithmes ne font pas que détecter et analyser anticipativement les opportunités et les risques : dans la mesure où ils permettent d'intervenir de façon « agile » sur le monde, ils transforment le monde en sélectionnant, parmi tous les « possibles », ceux qui sont les plus conformes à leur fonction objective.

On comprend dès-lors que la félicité des opérations algorithmiques ne se mesure à l'aune d'aucune « vérité » - en cela on ne peut pas à proprement parler, non plus, « d'erreur » algorithmique – mais seulement à l'aune de leur « opérationalité » : un algorithme doit produire « la » réponse opérationnelle et non équivoque au problème qui lui est posé, et ne pas tenir compte des autres réponses « équi-possibles », épistémologiquement ou empiriquement parlant, qui auraient pu être produites en dehors de la « fonction objectif » (c'est-à-dire des contraintes de coûts, de temps, de finalités) déterminées par son « donneur d'ordres » (le concepteur ou son client). Il va de soi que la réponse non-équivoque implique d'oblitérer la

[9] Les anglophones parlent de trade-off.

contingence du résultat, ce qui correspond alors précisément à la pratique – jugée frauduleuse dans le domaine de la publication scientifique – du « p-hacking » : opération par laquelle les chercheurs choisissent de ne rendre compte que des analyses qui produisent des résultats conformes à leurs attentes en omettant de rendre compte de la proportion probable de faux positifs ou de faux négatifs dans leurs résultats, c'est-à-dire de la dépendance de leurs résultats relativement à la méthode choisie.

L'impératif catégorique de l'optimisation algorithmique est qu'elle produise, chaque fois, une et une seule « réponse » ni nécessairement « vraie », ni nécessairement « fausse », mais opérationnelle, c'est-à-dire non équivoque, et suffisamment fiable pour dispenser de tout détour épistémologique ou herméneutique. Le décisionisme algorithmique (Parisi 2017) privilégie la rapidité et l'univocité de la décision sur la qualité de la décision (le fait qu'elle soit correcte), d'autant que dans un processus d'apprentissage continu, « l'erreur » n'est pas tant un défaut qu'une condition même de l'apprentissage. En régime décisioniste, le critère de félicité est le caractère « décisif ». Si l'on peut parler d'une forme de « dictature » des algorithmes c'est en cela, et en cela seulement – mais c'est énorme -, que la logique décisioniste s'émancipe d'un mode de résolution des problèmes qui fait assumer au décideur les conséquences de ses décisions, celles-ci ne lui étant plus imputables, dans la mesure où elles s'imposent sur le mode d'une forme de "nécessité", émancipée de toute justification, et apparaissent dès-lors comme revêtues de l'autorité immédiate et sui-generis des "choses en soi", libérées du jougs de la rationalité. L'action fondée sur la détection des émergences (emergency signifie aussi, en Anglais, urgence), se légitime précisément par ce qui, dans les rationalités gouvernementales libérales, limitait l'action des gouvernants : l'ignorance, la contingence, l'incertitude radicale.

Ces éléments de croyance liés à la super-intelligence et à l'autonomie des IA comme à l'objectivité et à la qualité de données qui pourraient être porteuses de valeur intrinsèque en reproduisant un réel irénique, correspondent à une forme de culture du numérique assez distincte de celle des humanités numériques, au sens d'un nouveau rapport à un milieu numérique dont les usages sont à inventer de manière créative (Zacklad 2020). D'un côté, la vision d'une technologie susceptible de dépasser les compétences humaines elles-mêmes totalement objectivables, de l'autre, la vision de technologies anthropologiquement constitutives de l'espèce humaine bien avant et bien après la révolution numérique, via des dispositifs de médiation en série intriquant toujours les dimensions matérielles et culturelles porteuses de subjectivité et vectrices de subjectivation. Une grande partie des autres controverses trouvent un éclairage dans cette opposition.

B. Controverse relative aux impacts sur l'emploi et aux transformations sociétales associées

Une des grandes angoisses générées par la troisième vague de l'IA, basée sur la supposée supériorité de son intelligence porte sur le caractère inéluctable des suppressions d'emploi, notamment dans le secteur tertiaire. Le plus représentatif de ces craintes a été le rapport de Frey et Osborne (2013) qui prévoyait 47% d'emplois menacés aux États-Unis. Ce rapport a été en partie contredit par le rapport du Conseil d'Orientation pour l'Emploi (2017) « Automatisation, numérisation et emploi » qui envisage lui 10% d'emplois menacés. Il utilise une méthode différente basée sur un indice d'automatisation de l'emploi agrégeant quatre définitions identifiées par la littérature économique « comme déterminant la vulnérabilité de l'emploi à l'automatisation dans les conditions technologiques actuelles » (Synthèse, p. 10).

Ce rapport, plus réaliste mais également pessimiste, considérait que les emplois caractérisés par le « manque de flexibilité », « la faible capacité d'adaptation », « la faible capacité à résoudre les problèmes », « l'absence d'interactions sociales » étaient particulièrement vulnérables. Le métier des agents d'entretiens correspondait ainsi au métier le plus susceptible d'être remplacé par l'automatisation. Dans un rapport d'étude (Zacklad 2018) nous avons argumenté en plaidant que cette fragilité n'était aucunement une fatalité.

En nous appuyant sur des travaux d'un consortium de recherche privé, le CRDIA[10], il est possible de montrer que le secteur du Facilities Management qui inclut l'accueil, le gardiennage, la maintenance, la propreté, etc. correspond à des besoins encore très largement insatisfaits et que la voie actuelle de la taylorisation, prolongée par l'automatisation, ne va pas dans le sens d'une réponse satisfaisante à ces besoins. Pour inverser la tendance, il faut développer des innovations servicielles basées sur la valeur « aménitaire » du service. Selon Baron et Cugier (2016), les aménités sont des aspects agréables de l'environnement ou de l'entourage social qui ne sont ni appropriables ni strictement quantifiables.

Le développement de prestations de service aménitaires, renvoie (1) à une analyse de la pertinence du service pour les bénéficiaires plutôt qu'à une vision de la qualité au sens industriel, (2) à des gains de productivité basés sur l'intégration services, (3) à la nécessité d'éviter les spécialisations par métiers excessives qui appauvrissent les tâches. Elle implique

[10] https://crdia.org/ (Consortium de Recherche et Développement de l'Ile Adam).

de valoriser les compétences des opérateurs sur certaines dimensions pointées dans l'étude du COE et notamment la capacité à résoudre des problèmes de manière autonome et l'accroissement des interactions sociales.

Les entreprises seraient donc libres de développer des stratégies d'innovation qui s'appuient sur trois leviers : une action sectorielle visant à inclure dans les marchés entre donneurs d'ordre et fournisseurs la valeur aménitaire, le développement des compétences des collaborateurs mais aussi des bénéficiaires, une nouvelle attitude du management qui évite la méthode descendante du « command and control » basée sur des indicateurs partiels. Ces stratégies relèvent de modèles de croissances alternatifs à l'industrialisation taylorienne intensive qui sont décrits dans différentes théories : économie de la fonctionnalité, économie territoriale ou de la proximité, économie de la singularité... mais aussi sur des visions anthropologiques qui remettent en cause la place excessive du marché et des sciences économiques dans l'analyse du travail et de la valeur, malgré leurs efforts pour élaborer des visions alternatives.

Cet exemple montre que les prévisions scientifiques relatives à la destruction de l'emploi que l'on pourrait simplement souhaiter compenser par des mesures d'accompagnement, peuvent être contredites radicalement par d'autres approches scientifiques dans le cadre d'une controverse éthique située. Les prévisions catastrophistes qui visent souvent à avoir un effet performatif (Muniesa & Callon, 2013) (prophéties auto-réalisatrices), peuvent soutenir de vieilles politiques de substitution du capital au travail, qui enrichissent une minorité d'actionnaires et d'institutions financières tout en produisant généralement des effets sociaux, environnementaux et sanitaires dramatiques. En utilisant d'autres stratégies d'innovation basées sur une approche « augmentative » du recours au numérique et à l'IA dans lesquels celui-ci accroit les compétences des opérateurs plutôt qu'ils ne les remplacent, il est même possible de développer l'emploi en pourvoyant à un grand nombre de besoins insatisfaits[11].

C. Controverses liées au travail et aux modalités d'organisation du travail

En lien direct avec les controverses liées à l'emploi, interviennent les controverses liées au travail et à son organisation mais également aux interactions de services de plus en plus automatisées. Depuis longtemps, les publications en management ou en ingénierie promouvant l'automatisation à outrance et vantant les bénéfices supposés d'une rationalisation intensive,

[11] Nous ne rentrons pas ici dans le débat consistant à s'interroger sur la pertinence du maintien du paradigme de l'emploi par rapport à d'autres formes de travail rémunéré qui pourraient apparaitre comme plus émancipatrices.

d'un contrôle par les données, de la quantification de l'activité des collaborateurs, sont remises en cause par des recherches en sciences humaines et sociales dans le domaine de la sociologie, de la psychologie ou des SIC ou par d'autres recherches en sciences de gestion d'inspiration plus qualitative.

Ces visions « optimisatrices » s'appuient généralement sur la croyance dans l'objectivité des données que nous avons déjà évoquée. Plus les opérateurs produiront de telles données plus il sera par ailleurs facile de développer des algorithmes d'IA renforçant le contrôle de leur activité. Ces visions se déploient dans tous les domaines, les ressources humaines, la formation professionnelle, la relation client, la logistique, etc. Elles s'appuient sur une vision de l'activité et du travail comme étant totalement réductibles à des procédures formalisées. Elles excluent la prise en compte des évènements imprévus, de la singularité, des rencontres, de l'incertitude, de l'innovation continue… Elles contredisent les travaux des ergonomes et des psychologues du travail qui opposent le travail réel au travail prescrit, ou qui insistent avec les SIC sur l'importance des collectifs de travail comme espace de délibération et de réinvention en continu de la performance et de ses modalités (Clot et Stimec 2013).

Au lieu d'envisager le numérique comme un outil au service de l'intelligence collective dans une logique augmentative (Zacklad 2012, 2020), elle l'envisage comme un outil de contrôle et de rationalisation visant in fine à pouvoir les télécommander de façon réactive. Soulignons que la rationalisation portée par l'IA et notamment la rationalisation connexionniste[12], diffère largement de la rationalisation traditionnelle. Comme nous l'avons développé plus haut, les algorithmes « apprenants » produisent des prescriptions qui peuvent être à la fois totalement personnalisées et non explicables. Elles sont jugées potentiellement « opérationnelles » du point de vue des critères de coûts, de temps ou de finalités déterminées par les concepteurs ou leurs donneurs d'ordre mais ne sont pas rationnelles au sens de l'explicitation des préférences et des valeurs qui justifieraient de manière contradictoire les décisions.

Cette vision est corollaire d'une autre approche dans le domaine de la relation client, consistant à penser que clients et usagers ont essentiellement besoin d'une accessibilité immédiate au service au détriment de relations interpersonnelles effectives. C'est le règne du déploiement

[12] C'est une rationalisation paradoxale puisqu'elle ne peut pas fournir la justification de ses prescriptions. C'est une forme d'optimisation dynamique détectant à un stade extrêmement précoce, des "émergences", des "propensions", des "opportunités" et qui, de ce fait, permettent à leurs "donneurs d'ordre" de capitaliser sur la contingence. L'exemple le plus parlant est le trading haute fréquence.

des chatbots et autres interfaces automatisées tentant de rivaliser avec les opérateurs humains en simulant des dialogues d'assistance de manière assez souvent peu pertinente. Ce sont aussi les travaux sur les robots dédiés à la prise en charge de la solitude des personnes âgées ou à celle des besoins sexuels insatisfaits, présentés comme le futur du « care ».

Or, là aussi, de nombreuses recherches en socio-économie des services remettent profondément en cause l'idée d'un bénéfice apporté par l'automatisation exclusive. Au contraire, pour ces chercheurs, une grande partie de la valeur du service repose sur des relations interpersonnelles visant à la transformation de l'état psychique, social, physique du bénéficiaire. Ces relations reposent sur la confiance interpersonnelle, l'identification à des professionnels ou à des collectifs, l'insertion dans des communautés. Elles sont souvent puissamment ancrées dans la culture et les territoires. L'automatisation intégrale du service est bien souvent une destruction de l'essentiel de ces services qui oriente vers des plateformes de plus en plus délocalisées et automatisées au bénéfice des grands monopoles du numérique et des opérations de standardisation qu'ils promeuvent. Dans cet espace de controverses également, il est possible d'adopter une position radicalement opposée à la taylorisation à outrance et à l'automatisation des services plutôt que de chercher à en atténuer les effets ou de chercher par tous les moyens à donner une apparence humaine aux robots.

D. Controverses liées à la citoyenneté, à la diversité, à l'égalité des chances et à la transparence

De nombreux acteurs mais aussi chercheurs souhaitent généraliser l'utilisation de la décision automatique dans les domaines du droit, de la médecine ou de l'orientation scolaire, de la gestion des ressources humaines, pour ne prendre que quelques exemples. Cependant les biais que nous avons évoqués plus haut posent des problèmes en matière juridique, de diversité et d'égalité des chances. Pour les chercheurs inscrits dans une logique plus « solutionniste », la résolution de ces problèmes implique de travailler sur la qualité des données qu'il faut rendre objectives (cf. supra) et sur la « transparence » des algorithmes qui renvoie essentiellement à des enjeux d'amélioration technique (cf. Besse et al. 2018).

Le terme de transparence possède un écho particulier dans le domaine de l'action publique. Comme le rappelle J. M. Sauvé (2011) l'action publique repose sur les deux faces d'un même dilemme éthique fondamental celui du respect du secret, notamment de la vie privée, et celui de la transparence, notamment celle des décisions administratives, des délibérations concernant le public, de la nature des données conservées sur les individus… Cette transparence se

prolonge par le principe, consacré par la loi pour une République numérique (2016), de transparence des algorithmes publics, notamment quand ils sont utilisés pour prendre des décisions administratives individuelles.

Malheureusement, les enjeux sous-jacents à cette exigence de transparence relèvent d'un faisceau de problèmes non susceptibles de trouver des solutions techniques.

- Les processus algorithmiques (inductifs, statistiques, indifférents à la causalité, impliquant souvent du « calcul en parallèle », métabolisant des données souvent peu structurées et peu denses en information…) ne se laissent pas aisément transcrire sous une forme linéaire compréhensible par les humains. C'est particulièrement vrai s'agissant des algorithmes de machine learning – spécialement lorsqu'ils sont peu ou pas supervisés – la « logique » de l'algorithme, qui « découvre » lui-même les variables à prendre en compte, qui affine lui-même ses « métriques » au fur et à mesure de son apprentissage - est particulièrement intraduisible sous une forme linéaire compréhensible par les humains. Pour les humains, l'IA connexionniste fonctionne suivant une « rationalité alien » (L. Parisi 2019).

- Par ailleurs, la transparence des algorithmes se heurte bien souvent au régime juridique de protection de la propriété intellectuelle et du secret industriel[13], ou à l'argument suivant lequel la transparence des algorithmes anéantirait leur efficacité : si l'on rendait publics les critères de détection des propensions terroristes, de fraudes, de récidives, de non-remboursement des dettes… il suffirait à ceux qui veulent échapper à la détection d'éviter d'envoyer les « mauvais signaux ».

- Enfin, même si le fonctionnement des algorithmes pouvait être en partie appréhendé par les « usagers », les résultats pourraient néanmoins présenter un caractère contre-intuitif, dans la mesure où les algorithmes sont incapables de comprendre le « contexte » dans lequel on les fait intervenir, la décision pouvant sembler « sortie de nulle part ». Ce caractère contre-intuitif peut rejoindre la problématique du manque de

[13] cf. la mésaventure de la Task Force instaurée par le New York City Council pour assurer la transparence des algorithmes impliqués dans les politiques publiques… https://www.citylab.com/equity/2019/12/ai-technology-computer-algorithm-cities-automated-systems/603349/

justification des résultats, c'est-à-dire de leur manque d'adéquation avec ce que l'on peut considérer comme juste, raisonnable, équitable, pertinent… Selon les termes de cette controverse éthique, est remise en cause la possibilité même d'une généralisation de la décision automatique et autonome concernant le public dans les domaines que nous avons évoqués, comme l'exonération de la responsabilité des institutions en charge de ces questions. Il faut donc toujours in fine, s'en remettre à l'avis d'un professionnel seul capable de contextualiser et de pondérer les propositions des algorithmes en fonction du cas singulier.

Le mouvement « fair, accountable, transparent machine learning » tente de trouver des garanties techniques contre les risques de discrimination, d'opacité des décisions et de non-imputabilité dans les « décisions » algorithmiques. Aussi bien intentionné que soit ce mouvement, il accrédite une vision extrêmement réductrice 1) de ce en quoi peut consister la justice (qui serait réductible à certaines propriétés des « boîtes noires » algorithmiques, c'est-à-dire de sous-systèmes techniques, alors qu'elle est un principe de perfectibilité du « système » social dans son ensemble) ; 2) de ce en quoi peut consister la responsabilité (réduite aux obligations bureaucratiques de « compliance » avec les exigences du RGPD, en Europe) ; 3) et de ce en quoi peut consister l'exigence de publicité (du caractère « public » et politique de la délibération relative aux critères de mérite, de besoin, de désirabilité, de dangerosité, qui président, dans nos sociétés, à la répartition des ressource et des opportunités), disqualifiée au profit de la « transparence » de la décision individuelle clôturée sur elle-même rendue non rapportable aux enjeux structurels, collectifs.

E. Controverses liées aux modalités de participation des usagers (et non usagers) aux choix techniques qui les impactent

Un des problèmes classiques du déploiement des nouvelles technologies sur la base de leurs supposés avantages indiscutables, ici l'intelligence des algorithmes supérieure à celle des humains dont la généralisation à marche forcée est censément bénéfique pour les organisations et l'économie, est que la conception de ces algorithmes et de leurs modalités d'usage est faite par des concepteurs de manière le plus souvent descendante. Pour reprendre les termes de Callon, Lascoumes et Barthe (Callon et al. 2014), la conception se déroule « en laboratoire » sans aucune prise en compte de l'expérience des « chercheurs de plein air », les usagers ou des personnes qui pourraient être impactées par ces transformations sur le terrain.

Or, pour de nombreux praticiens et chercheurs, les transformations numériques ne peuvent plus être mises en place sans recourir à des méthodes de co-design qui impliquent de manière directe les utilisateurs et les parties-prenantes (Foliot et al. 2020, Zacklad 2020). Le « design de solution », qui vise à concevoir les artefacts, n'est pleinement réussi que s'il s'accompagne d'un « design de relation » (Zacklad et al. 2021) qui contribue à définir un public, au sens où nous l'avons défini plus haut, mais aussi à s'assurer que ce public puisse intervenir activement dans la spécification et la gouvernance des technologies déployées.

Dans la sphère des services publics, la transition numérique conduite sans « design social » conduit à exclure un nombre toujours plus grand de non-usagers du numérique. Cela peut concerner les enjeux de l'accès physique, comme la suppression des cabines téléphoniques, mais cela concerne également le conseil, avec la fermeture des guichets physiques des administrations et peut-être à terme le service téléphonique rendu par des conseiller humains, des décisions qui seraient catastrophiques vu l'incapacité à contextualiser des algorithmes connexionnistes que nous avons déjà évoquée.

F. Controverses liées à la neutralité carbone du numérique et à l'aménagement du territoire

La sixième controverse concerne les enjeux environnementaux et porte sur les avantages supposés du numérique en termes de neutralité carbone mais aussi de mobilité et d'aménagement du territoire. En supprimant les déplacements physiques, le numérique aurait un effet bénéfique. Mais il est maintenant établi que la consommation énergétique du numérique, de la production des équipements aux data center, est considérable (Flipot et al. 2013). L'IA est particulièrement gourmande du fait de sa consommation de données massives, notamment dans le domaine du langage naturel[14].

Mais, par ailleurs, il y a une tension entre les arguments de ceux qui soulignent la propension du numérique à faciliter le télétravail, et donc une certaine délocalisation des emplois, et ceux qui soulignent les avantages de son usage dans le déploiement de téléservices permettant la délocalisation des services publics et privés qui entraine des effets de bord néfastes en matière d'aménagement du territoire, mais aussi potentiellement relationnels si le télétravail n'est pas pensé. La désertification des territoires en termes d'emplois locaux, entraîne d'autres formes

[14] https://www.cnetfrance.fr/news/pourquoi-l-intelligence-artificielle-est-un-desastre-ecologique-39886927.htm

de mobilité contraintes et appauvrit le tissu local sur un plan économique et culturel nuisant à la durabilité.

Les territoires produisant moins de données sont de plus en invisibles et leur invisibilité renforce encore la centralisation. L'utilisation raisonnée du numérique doit contribuer à ce que local soit toujours et partout producteur de ressources et de richesses sur les plans énergétique, agricole, industriel, serviciel et symbolique, comme en témoigne le mouvement des villes en transition (Hopkins et al. 2017) l'accroissement de la diversité étant sans doute le meilleur rempart contre les risques d'effondrement écologique en matière d'espèces vivantes comme de pratiques culturelles, les deux étant fortement interdépendants.

La soutenabilité écologique doit s'entendre à la fois comme soutenabilité environnementale, comme soutenabilité sociale et comme soutenabilité psychique. A défaut de prendre en compte ces trois dimensions simultanément, on débouche sur des affirmations comme celles-ci : le télétravail est bon pour l'écologie alors que, mal géré, il peut être néfaste pour l'écologie sociale et psychique. Les enjeux sociaux, psychiques et environnementaux sont indissociables, la seule voie possible étant de percevoir qu'aucune justice environnementale n'est atteignable sans garantir en même temps la justice sociale et l'écologie psychique (cf. Guattari,1989).

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 qui justifient le recours à la technologie en ouvrant des espaces de controverses scientifiques masquées, telles des boites noires, par 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[15] ou d'ouverture vers d'autres possibles.

Car nous ne pensons pas que le rôle 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 la 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 de la démocratie technique (Callon et al. 2014), des sciences citoyennes ou, pour le dire dans les termes de Stengers, à 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

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

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.

[15] 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.

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

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

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

Deleuze, G. (1968). *Différence et répétition*. PUF.

Desrosières, A. (2014). *Prouver et gouverner. Une analyse politique des statistiques publiques*. La Découverte, p.58.

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

Dewey, J. (1941), "Propositions, Warranted Assertibility, and Truth", The Journal of Philosophy, 38(7): 169–186.

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

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

Foliot, C., Serikoff, G., Zacklad, M. (2020). *Le Lab des Labs*. https://hal.archives-ouvertes.fr/hal-02437318

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

Geman, S., Bienenstock, E., and Doursat, R. (1992) "Neural networks and the bias/variance dilemma". Neural Computation, 4(1):1–58.

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/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

Land, N. (2014), “The teleological identity of capitalism and artificial intelligence”, Remarks to the participants of the Incredible Machines. Digitality and Modern Systems of Knowledge at the Threshold of the 21st Century, Goldcorp Centre for the Arts,Vancouver, March 8, 2014.

Latour B. 2007. « Pensée retenue, Pensée distribuée », in Jacob C., *Les lieux de savoir. Espaces et communautés*, p. 605-615.

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/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

Prairat, E. (2014). Valuation et évaluation dans la pensée de Dewey. *Le Telemaque*, *n° 46*(2), 167-176.

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, (2016b), “Algorithmic governmentality: radicalisation and immune strategy of capitalism and neoliberalism?”, La Deleuziana, n.30.

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.

Sauvé, J.-M. (2011). *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

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. (2018). *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. (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

Zacklad, M., Arruabarrena, B., Berthinier-Poncet, A., & Guezel, N. (2021). Les labs d'innovation interne : Typologie des innovations, approche plateforme, rôle du design. *Approches Théoriques en Information-Communication (ATIC)*, 2(1), 127-161. Cairn.info. https://doi.org/10.3917/atic.002.0127

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