Vers une éthique processuelle de l'IA

Eric Pardoux, Louis Devillaine

Vers une éthique processuelle de l'IA

Éric Pardoux[1,2], Louis Devillaine[3]

[1] CNRS, IHRIM (UMR 5317), ENS Lyon
[2] CNRS, Maison Française d'Oxford (UMIFRE)
[3] Univ. Grenoble Alpes, CNRS, Sciences Po Grenoble, Pacte, 38000 Grenoble, France

eric.pardoux[at]ens-lyon.fr / louis.devillaine[at]univ-grenoble-alpes.fr

Résumé

Nous mettons en évidence certaines des dimensions éthiques présentes dans le processus de design *des systèmes d'AM. Sur cette base, nous proposons une éthique processuelle qui associe des enjeux éthiques à toutes les étapes de la conception. Cette démarche nous amène à mettre en lumière différents instants du processus de* design *qui peuvent occasionner des préoccupations d'ordre éthique. Il s'agit de présenter la base de réflexions futures sur la place que l'éthique peut prendre dans la production d'IA en général.*

Mots-clés

Éthique de l'IA, Éthique processuelle, Éthique by design.

Abstract

We highlight some of the ethical dimensions present in the design of ML systems. Based on this, we propose to rethink ethics as process undergoing at all stages of the design process. This approach leads us to underline different moments in the design process that may give rise to ethical concerns. Our aim is to present the basis for future reflections on the place that ethics can take in the production of AI.

Keywords

AI ethics, Processual ethics, Ethics by design.

1 Introduction

Il n'est probablement plus réellement nécessaire de démontrer les avancées grandissantes de l'intelligence artificielle (IA) ces dernières années [31]. Ces avancées ne sont pas sans soulever des questions éthiques [1] liées aux dommages que l'introduction de l'IA pourrait provoquer : reproduction de biais sociaux, discrimination, dangers de l'automation sont quelques-uns des enjeux éthiques que l'on peut citer [35, 41, 51]. Lors de son déploiement, l'IA peut ainsi venir perturber des valeurs morales et sociales comme la justice, l'équité ou la responsabilité [2]. Notre objectif ici est d'introduire une réflexion sur la façon dont la considération des enjeux éthiques de l'IA peut avoir lieu tout au long de son développement. Pour ce faire nous proposerons l'ébauche d'un cadre éthique qui permettra d'appréhender l'éthique de l'IA au travers de son processus de *design* [3]. Notre ambition est notamment de proposer une approche complémentaire à l'éthique par principes qui nous semble dominer la littérature de l'éthique de l'IA [40]. Nous suggérerons ainsi l'adoption accrue d'une éthique processuelle, accompagnant le *design* des systèmes d'IA (SIA) de leur idée initiale à leurs usages finaux. Comme nous le verrons, cette éthique processuelle de l'IA doit se traduire par une prise de conscience plus marquée des enjeux éthiques souvent implicites et pourtant cruciaux présents tout au long du processus de conception des SIA.

Avant de rentrer dans le détail de ces considérations, il convient de préciser à quoi nous nous référons lorsque nous employons le terme d'IA. Nous ancrons avant tout nos considérations à partir des algorithmes basés sur l'apprentissage machine (AM) [4]. Nous faisons ce choix car il permet d'aborder des applications émergentes de l'IA, qui posent des questions nouvelles à l'éthique — par exemple le problème de l'explicabilité, ou la responsabilité pour les SIA à apprentissage incrémental. De plus, nous considérons que certains des lieux de l'éthique présentés pour les algorithmes d'AM sont communs avec le développement de SIA non basés sur l'AM. En cela, nous espérons fournir

1. Par éthique nous entendons ici la façon dont des principes moraux et des valeurs peuvent se traduire dans les actions et les décisions, notamment dans le développement et l'introduction de nouvelles technologies. Il existe une pluralité de perspectives éthiques — par exemple le conséquentialisme, l'approche déontologique ou encore l'éthique des vertus. Au-delà de leurs différences d'implémentation, leur principal point commun réside dans la fixation d'un horizon d'attente par rapport à ce que le réel devrait être (d'un point de vue moral, social ou encore politique). Cela reste le cas, que le jugement éthique des actions et décisions se traduise (i) dans la considération de leurs conséquences, (ii) des principes suivis pour les réaliser ou (iii) au travers des vertus mobilisées par l'agent moral dans leur réalisation.

2. Il convient toutefois de souligner que ces perturbations peuvent être bénéfiques si elles promeuvent et améliorent la qualité morale des actions et décisions auxquelles elles contribuent.

3. Nous employons le terme de *design* dans son acception anglaise, afin de recouvrir les différents sens qu'on peut lui donner. Ceux-ci seront explicités dans la suite de l'article.

4. Pour expliciter notre compréhension, nous repartons de la définition donnée par T. Mitchell [36] : « on dit d'un programme informatique qu'il apprend d'une expérience E vis-à-vis d'une classe de tâches T et d'une mesure de performance P si ses performances sur les tâches T, mesurées par P, s'améliorent avec l'expérience E ». En l'occurrence, l'expérience peut consister en des exemples — on parle alors d'apprentissage supervisé, ou non-supervisé — ou bien en un ensemble de règles — apprentissage par renforcement.

des réflexions utiles à un cadre plus large qui dépasserait les limites seules de l'AM.
Cet article sera organisé en deux temps principaux : tout d'abord nous formulerons une proposition de reconstruction d'un cadre éthique dans lequel inscrire les pratiques de *design* d'AM. Nous aborderons ensuite plus en détail comment cette posture éthique théorique pourrait s'illustrer dans le cadre du développement de systèmes d'AM (SAM) [5]. Pour ce faire, nous décrirons le processus de *design* de SAM pour faire émerger les lieux des enjeux éthiques en son sein.

2 Où est l'éthique de l'IA ?

2.1 Une exigence éthique sur la technique ?

Les préoccupations éthiques entourant le développement de l'IA engendrent une volonté de régulation qui ne concerne plus seulement les usages. Dans les applications dites à « haut risque » telles que les applications en santé, les cadres éthiques prescrivent tout autant les usages qui devraient être faits des SIA que ce qu'ils devraient être en eux-mêmes. Cela se traduit par exemple par une exigence de certaines caractéristiques : les SIA doivent se révéler être explicables, transparents, justes, équitables, non discriminatoires et tout ceci dans leur conception même [6]. On retrouve ces exigences dans la littérature en éthique de l'IA tout comme dans le cadre de régulations [25].
À l'échelon européen, de nombreuses propositions sont formulées ces dernières années pour parvenir à mettre en place un cadre régulateur commun au développement des SIA à « haut risque » [7]. La dernière proposition en date contient ainsi un article prônant une exigence de transparence [8].
Au niveau français, l'article 17 de la loi relative à la bioéthique [9] demande par exemple à ce que les professionnels de santé utilisant un SAM « s'assure[nt] que la personne concernée en a été informée et qu'elle est, le cas échéant, avertie de l'interprétation qui en résulte. ». Dans le même article est également inscrite une exigence pour que les « concepteurs d'un traitement algorithmique [...] s'assurent de l'explicabilité de son fonctionnement pour les utilisateurs ».
Dans un cadre comme dans l'autre, il est intéressant de remarquer l'absence de régulation forte des usages de l'IA, au profit d'un conditionnement technique avant tout. En effet, certaines applications sont certes proscrites [10], mais aucune mention n'est faite d'un encadrement strict des usages une fois que les SAM sont autorisés et certifiés, ni d'un suivi du système une fois déployé.
La direction prise par la formulation juridique de ces exigences techniques, tout comme la perspective employée dans une large part de la littérature sur l'éthique de l'IA relève selon nous d'une certaine posture éthique. Cette dernière reviendrait à concevoir l'objet et son éthique à partir de son usage — notamment pour déterminer s'il est à « haut risque » ou non — mais agirait avant tout en conditionnant la conception et les caractéristiques techniques des SIA. Une telle pratique pourrait s'interpréter comme renvoyant à une « éthique *by design* » [11].
Dans la suite de cet article, nous nous proposons de présenter les enjeux de cette démarche d'éthique *by design*, avant de faire le lien avec le processus de *design* d'IA. Cela nous permettra de mettre en avant les étapes au cours desquelles il peut être bénéfique de mettre en place un questionnement éthique lors du *design* de SAM.

2.2 L'éthique *by design*

L'éthique par *design* est un concept relativement récent. Selon F. Fischer [13], la première occurrence académique du terme d'« *ethics by design* » date de 2010, dans le contexte de la stratégie d'entreprise [39]. Il s'agit alors d'une ouverture des entreprises vers une éthique conséquentialiste. Cette démarche consiste à ne plus seulement constater les changements opérés par les pratiques managériales ou l'introduction de technologies dans un contexte organisationnel, mais à les anticiper pour en prévenir les maux potentiels. Elle dénote ainsi d'une volonté d'opérer une éthique *a priori* par rapport à l'usage, qui soit en anticipation des éventuelles nuisances liées à l'introduction d'une nouvelle technique. Il ne s'agit donc pas seulement de mettre en place des pratiques vertueuses, qui relèveraient d'une déontologie professionnelle au sein des entreprises mais bien de prendre en compte l'intégralité des processus agissant au sein de l'organisation et des conséquences en découlant. Cette première occurrence de l'éthique *by design*, issue de la littérature du management, destine donc la réflexion éthique à toutes les étapes des processus d'organisation et de décision qui peuvent être menés dans les entreprises.
Dans le contexte francophone, le terme de *design* est souvent repris directement de l'anglais : il nous semble important de souligner la diversité des sens dont il peut relever, afin de mieux problématiser les pratiques relevant de l'éthique *by design*. Trois visions principales de l'éthique *by design* peuvent être adoptées par les *designers* en fonction du sens donné au *design* [12].

Une éthique par dessein. Elle peut être entendue en premier lieu dans sa traduction littérale comme une éthique par

5. Par système d'apprentissage machine, nous entendons ici un ensemble hétéroclite d'objets techniques, incluant aussi bien les programmes informatiques qui implémentent les méthodes d'AM que les interfaces humain-machine (IHM) et les systèmes sociotechniques qui les intègrent.

6. Néanmoins, le lien entre ces caractéristiques « éthiques » et des propriétés techniques intrinsèques aux algorithmes n'est pas une évidence. Sur le cas de l'équité voir par exemple l'étude menée par S. Wachter *et al.* [51]. Sur les obstacles techniques à l'élaboration d'un système explicable, se référer à la revue de Dazeley *et al.* [9]

7. Dans cette catégorie large, sont incluses notamment les SIA pour la santé, les systèmes destinés à la justice, à l'éducation, aux prestations sociales, aux moyens de répression, etc. Voir l'annexe III de la *Proposition de règlement du Parlement européen et du Conseil Établissant des règles harmonisées concernant l'IA* (COM/2021/206 final) parue le 21/04/2021.

8. Voir l'article 13 de la proposition citée précédemment.

9. Loi n° 2021-1017 du 2 août 2021 relative à la bioéthique (1), NOR : SSAX1917211L.

10. Pour le règlement européen, il s'agit notamment des SIA ayant recours à des « techniques subliminales » de manipulation, à l'exploitation de vulnérabilités, à la notation sociale ou encore à l'identification biométrique à distance en temps réel.

11. Notre constat semble par ailleurs appuyé par l'existence d'un document de travail de l'UE portant sur des recommandations suivant une certaine acception de l'éthique *by design* pour les chercheurs en IA [8].

dessein, l'intention des *designers* étant supposée se traduire dans le dispositif technique de façon positive. Le dispositif technique est conçu pour être déployé dans un certain cadre où il serait éthique par nature, du fait de l'intention transmise par les *designers* [12]. Les futurs usagers seraient empêchés de produire des effets indésirables car ils suivraient cet usage prescrit par intention. Dans une approche d'AM incrémental et non statique, on peut voir les limites d'une telle conception de l'éthique par dessein au travers de l'exemple de Tay, l'agent conversationnel déployé par Microsoft sur Twitter. L'usage détourné du système a dépassé les limites anticipées par les développeurs jusqu'à l'apprentissage d'injures [52]. Le dessein des *designers* s'est ainsi retrouvé débordé par l'usage, montrant les limites de l'absence de contraintes techniques.

Une éthique par définition. L'éthique *by design* s'envisage ici sous le prisme de la conception même de l'objet technique, qui est supposé incarner des valeurs désirables. Son usage dans la société serait maîtrisé par les *designers* au travers de la définition de ses spécifications. C'est l'agencement même des composants techniques qui vient alors conditionner les actes éthiques de ces derniers, pour reprendre la formulation de Fischer, l'éthique devient ainsi « technicisée » [12, p. 64]. On retrouve ici notamment les approches de *Privacy by design* [45], reprises par Fischer. Ceci peut être considéré comme une éthique par définition, car les fonctionnalités de protection de la vie privée sont alors intégrées dès le début du processus de *design* et non pas en seconde intention une fois le cœur du système technique déjà développé. Cette approche est critiquable du fait de la potentielle rigidité du système une fois déployé, qui peut conduire à un rejet de la part des usagers, si les contraintes pour respecter la vie privée deviennent plus fortes que les bénéfices qu'elles fournissent.

Une éthique de la médiation. Une dernière proposition est l'éthique *by design* fondée sur une approche relationnelle entre usager et objet, en incorporant une réflexion sur les conséquences probables de l'interaction du système avec les individus constituant la société. Cette approche intègre en quelque sorte les deux dimensions précédentes : la technique est une médiation entre l'humain et le monde, elle conditionne sa vision de ce dernier et les actions qu'il peut entreprendre à son propos. L'éthique n'est plus centrée sur l'humain mais sur ses relations à son milieu — aussi bien technique que social. En ce sens, il nous paraît bénéfique de favoriser cette éthique de la médiation qui enjoint à systématiquement contextualiser socialement l'usage qui sera fait des techniques développées. Il s'agit donc d'avoir des pratiques du *design* qui ne soient ni trop ouvertes — comme dans une éthique par dessein — ni trop fermée — comme dans des systèmes trop contraints conçus par définition. L'idée est de laisser à l'usager la main sur la paramétrisation des systèmes en prenant en compte les potentiels usages qu'il en aura [28]. Cette approche plus équilibrée permettrait de diriger les usagers vers un questionnement ou un comportement éthique souhaitable, sans pour autant être trop paternaliste [14]. C'est une façon d'échapper à l'impasse d'une relativisation totale des enjeux éthiques de la technique par la prétendue infinité de ses usages et détournements possibles. L'approche relationnelle promue par la médiation nous paraît transcrire au mieux la réalité des processus de *design*, de réception et d'intégration des systèmes sociotechniques dans le monde réel.

Ces trois approches sont évidemment non exclusives — dans le sens où l'éthique de la médiation peut mobiliser des approches par définition ou par dessein. Elles donnent à voir un panorama étoffé des différentes façons d'aborder la conception d'un système sociotechnique [13].

2.3 Ce que fait et peut le *design*

Il convient ici à notre sens de nous arrêter un temps sur ce à quoi renvoie l'idée de *design*, dans un cadre général. Comme la typologie de l'éthique *by design* suggérée par Fischer nous le fait entrevoir, ce terme est associé à une diversité de significations et de pratiques. Une première nuance forte est à faire avec la dimension dynamique du *design*, qui est alors analogue à la conception de l'objet technique : le *design* est un processus, un enchaînement d'actions et de décisions qui mène à la création d'un dispositif technique.

Pour autant, le *design* peut également renvoyer au résultat en tant que tel de ce processus. C'est alors le sens d'agencement sociotechnique qui prime avant tout. De par cet agencement particulier à tout dispositif, comme on l'a vu auparavant, certaines actions peuvent être favorisées ou au contraire entravées voire rendues impossibles. Nous retrouvons ici l'hypothèse de L. Winner [53], selon laquelle la dimension politique et éthique des objets découle avant tout de leur organisation et de leurs interrelations plutôt que de leurs caractéristiques intrinsèques. Si la littérature philosophique est divisée sur la façon avec laquelle émergent ces valeurs politiques [14], il reste un constat fort : la technique a notamment pour fonction de modifier l'état du monde et le processus de *design* y participe directement.

En effet, ce dernier implique de chercher à transformer le monde en ce qu'il *devrait être* [15]. Le *design* est donc tout autant un processus de composition (d'agencement) d'un nouveau réel qu'un processus de recomposition (réagencement) d'un état passé du monde et du tissu sociotechnique. En cela encore, il possède certaines similitudes avec le processus d'innovation [19]. D'un part, parce que le *design*

12. Ici et par la suite nous utiliserons ce terme dans un sens large incluant toute personne contribuant de près ou de loin au processus de *design* : cela va des commanditaires aux ingénieures, en passant par les programmeurs ou encore les chercheuses.

13. C'est ici une vision que nous tenons à défendre : les dispositifs techniques ne font sens qu'une fois déployés dans un contexte social donné, coproduisant alors une nouvelle réalité aux dimensions à la fois techniques, sociales, éthiques, juridiques ou encore politiques [24].

14. Citons par exemple, en réponse à Winner, la critique de S. Woolgar et G. Cooper [55] ou encore la reconstruction plus récente de S. Lavelle qui développe cette réflexion à propos de la société numérique [29].

15. « *The engineer, and more generally the designer, is concerned with how things* ought *to be — how they ought to be in order to* attain goals*, and to* function. » (« L'ingénieur, et plus généralement le *designer*, se préoccupe de ce que les choses *devraient* être — comment elles devraient être dans le but d'*atteindre des objectifs*, et de *fonctionner* » [48, pp. 4-5 (notre traduction)]. Kraemer *et al.* proposent une définition similaire dans le cadre de l'éthique des algorithmes [28].

possède une dualité : il est un processus tout en étant le résultat de ce même processus. D'autre part, il est également une activité de réorganisation du réel [16] au travers de dispositifs techniques.

Au-delà de cette fonction de transformation, le *design* s'incarne dans une grande pluralité de pratiques. L'identification et la classification de ces dernières peut s'articuler autour de la place laissée à l'usager de l'objet dans la pratique du *design*. Nous insistons sur l'importance de distinguer les usages — qui renvoient aux pratiques réelles et situées d'un système — de l'utilisation — qui est une spécification issue du *design*, indiquant un ensemble restreint d'emplois, le plus souvent acontextuels [17]. Penser par les usages possibles, c'est donc anticiper détournements et dévoiements et par là même les soucis éthiques qui pourraient découler entre l'utilisation (projetée) et l'usage (effectif).

Trois grandes approches du *design* coexistent dans cette approche centrée sur l'usage au lieu de l'utilisation [18] : le *design* centré sur l'usager, la démarche de *design* participatif et enfin le *metadesign*. Les deux premières approches sont voisines. Dans le *design* centré sur l'usager, le *designer* projette les usages à venir de son dispositif sur un usager (fictif ou non) dont il essaie d'anticiper les réactions. L'ambition est alors de favoriser certaines actions ou d'en proscrire d'autres de par l'agencement du dispositif. Le *design* participatif (ou co-*design*) suit une démarche similaire mais en intégrant directement au processus de *design* un ou plusieurs usagers du futur dispositif [19].

Le *metadesign*, se détache des deux approches précédentes en prescrivant de façon moins restreinte les usages du dispositif. Il s'agit alors de laisser une plus grande marge de manœuvre à l'usager pour qu'il soit moins contraint par l'usage prescrit, en lui laissant une liberté de création et d'initiative au sein du système sociotechnique déployé. Cette dernière approche est particulièrement intéressante dans le champ du développement logiciel, qui connaît des évolutions fréquentes et une capacité de portage élevée. Pour attrayante que soit cette notion de *metadesign*, il convient de rappeler que le concepteur initial du système conserve tout de même une emprise forte sur les potentialités proposées de même qu'un regard sur les activités qui utilisent son dispositif. Toutefois, on marque là la possibilité d'une nuance forte entre un dispositif technique qui serait implémenté de façon rigide et donnant peu d'emprise à l'usager et une approche laissant les coudées plus franches à ce dernier.

Cette nuance entre les différents types de démarches de *design* par l'usage souligne plusieurs éléments importants : l'évaluation éthique dépend des usages possibles, et l'éventail de ces derniers dépend en partie du degré d'ouverture ou de fermeture du système technique. L'intention de transformation du réel qui dirige le processus de *design* peut être partagée entre différents acteurs, qui n'ont pas les mêmes compétences vis-à-vis du processus de conception, ni vis-à-vis de l'usage qui sera fait de son résultat. Il est à noter par ailleurs que ces trois approches ne sont en rien exclusives : des pratiques de *design* participatif où différentes parties prenantes sont à l'œuvre peuvent concevoir un système suivant les principes du *metadesign*.

Si le terme de *design* inclut ainsi en lui-même la volonté de transformer le monde selon une certaine intention, qu'apporte donc la notion d'éthique *by design* ?

Selon nous c'est la posture réflexive qui vient donner toute sa nuance et sa puissance à l'éthique *by design*. L'intention du *designer* porte consciemment ou non des enjeux éthiques qu'il faut interroger tout au long du processus. En s'intéressant à l'éthique par le prisme du *design*, nous suggérons qu'il est nécessaire de clarifier les enjeux éthiques présents tout au long des étapes de conception et de production des dispositifs techniques. Chacun des choix effectués dans le processus de *design* peuvent se trouver répercutés dans l'usage effectif du dispositif finalement produit. Comme nous l'avons déjà souligné, ce processus peut impliquer une pluralité importante d'acteurs : se poser la question de ceux qui peuvent avoir une posture de partie prenante en son sein, c'est aussi pratiquer un choix éthique [2]. Défendre un tel processus de mise en question à la fois technique mais également sociale du *design*, c'est défendre la possibilité de construire un SIA qui puisse se montrer réellement digne de confiance [20]. Nous allons voir par la suite comment ces enjeux peuvent se traduire plus particulièrement dans le cadre du développement d'un SAM.

3 Une éthique tout au long du *design*

3.1 Pour une éthique processuelle

3.1.1 Procédure ou processus ? Repenser l'éthique

Il s'agit selon nous avant de tout de faire évoluer la façon dont les enjeux éthiques sont considérés durant le *design* de SAM. Par une posture réflexive adoptée tout au long du processus de *design*, nous entendons ainsi plaider en faveur d'un regard critique sur les activités et choix de conception entourant ces systèmes.

Repenser la technique de façon critique implique de décentrer la focale du dispositif technique seul pour l'élargir au tissu social dans lequel il sera intégré. Plutôt qu'une remise en cause frontale des valeurs éthiques à adopter, il s'agit ici de refonder la façon dont nous les mettons en œuvre. Cela revient à ne plus se contenter uniquement de soulever ponctuellement les questions éthiques, lors de la certification ou

16. À la suite de V. Flusser, il est intéressant de noter que le *designer* peut être vu comme un « comploteur rusé qui tend ses pièges » (« A designer is a cunning plotter laying his traps. » (notre traduction)) au travers de ce qu'il conçoit [16, p. 17]

17. Autrement dit, l'utilisation c'est l'usage idéalement prescrit par les *designers*, ce qui nous renvoie à un *design* uniquement par dessein ou définition, sans tenir compte des effets de médiation.

18. Nous reprenons cette distinction depuis les travaux de V. Beaubois, présentés lors de la conférence « L'invention de l'usager » le 17/02/2022 et à paraître avant fin 2022.

19. De manière intéressante, cette démarche peut également être étendue en intégrant des professionnels de l'éthique au processus de *design* [50].

20. Ici nous souhaitons mettre l'emphase sur le fait qu'il s'agit d'une confiance à construire dans l'intégralité de la chaîne du *design* [27]. L'IHM n'est qu'une incarnation superficielle du *design* global qui traverse le système de ses caractéristiques techniques jusqu'à ses modalités d'intégration au contexte d'usage.

du dépôt d'un projet de recherche par exemple.
Comme Perrin *et al.* le soulignent, une telle pratique ponctuelle de l'éthique peut faire craindre de voir toute réflexion éthique « vidé[e] de son contenu pour devenir une simple procédure administrative, servant avant tout à protéger les institutions et les chercheur·euse·s » [44, p. 237]. Au contraire nous défendons une vision processuelle de l'éthique qui compléterait l'éthique procédurale [21]. Il convient pour cela de distinguer les procédures des processus. Les procédures sont des dispositifs mobilisés pour accomplir des fonctions prédéfinies, elles sont donc de nature fermées, discrètes, statiques et acontextuelles — pour une procédure donnée, sa fonction est accomplie quelle que soit la situation. En contraste, les processus nécessitent une certaine disposition face à une situation donnée : cela implique une capacité réflexive qui les rend ouverts, continus, dynamiques et adaptatifs. C'est sur la base de cette distinction que nous défendons l'intégration plus forte d'une dimension processuelle à l'éthique des *designers*. En percevant l'éthique comme un processus et non comme une simple procédure à accomplir ponctuellement, on évite de réduire l'éthique à une simple condition à remplir, une étape à satisfaire qui serait toujours identique, quel que soit le contexte.
Selon Perrin *et al.*, l'éthique procédurale est emblématique des sciences biomédicales tandis que son pendant processuel s'inspire grandement de l'approche pratiquée en anthropologie et dans les sciences sociales qualitatives [44]. Dans le cadre de la recherche, l'éthique procédurale se réalise avant tout par l'accomplissement de procédures, par exemple dans un dispositif administratif encadrant la recherche ou le développement d'une technique. Elle se traduit par exemple au travers de l'accent mis sur l'obtention du consentement éclairé des usagers ou des sujets de recherche, le passage devant des comités d'éthique ou l'élaboration de dossiers éthiques aux critères stricts et universels. En cela, elle permet tout de même de fixer un cadre favorable à la protection des individus observés. Au contraire, l'éthique processuelle affirme la dimension politique et donc éthique de chaque étape de la recherche [30] : les anthropologues doivent arbitrer leurs décisions en fonction des différentes parties prenantes de leur recherche.
Ce détour par l'anthropologie n'est pas anodin : comme le suggèrent déjà M. C. Elish et d. boyd, l'AM peut se concevoir comme une « ethnographie computationnelle » [11] [22]. Reconnaître le caractère non neutre de la technique implique une prise de conscience des responsabilités portées par les *designers* vis-à-vis du terrain dans lequel le dispositif devra s'intégrer. Cela implique une compréhension — quasi ethnographique donc — de ce contexte de déploiement. Ce rapport au terrain peut passer par le questionnement constant porté par l'éthique processuelle. Cette dernière se montre ainsi complémentaire à une éthique plus ancrée dans des procédures administratives. L'essentiel de la nuance entre procédure et processus réside dans une approche différente des temporalités de l'éthique [23] : d'une part l'éthique procédurale est discrète, figée dans certaines procédures, d'autre part l'éthique processuelle est continue, tout au long de la recherche ou du développement. L'idée de processus se traduit au travers de dispositifs différents (comme les conférences de citoyens ou les *focus groups* par exemple), plus adaptés aux contextes particuliers, dont l'existence peut venir en soutien des procédures éthiques plus classiques pré-existantes. Cette hybridation entre deux manières d'envisager l'éthique nous paraît être une piste intéressante pour résoudre les tensions entre la visée universelle d'une éthique basée sur des grands principes et le caractère exceptionnel que peuvent revêtir certaines techniques liées à l'AM [2]. La mise en avant de la possibilité d'une éthique processuelle ne doit ainsi pas effacer les effets bénéfiques que les procédures éthiques peuvent avoir pour protéger les futurs usagers d'une technique. L'exercice continu d'un raisonnement éthique tout au long du *design* permet d'intégrer une plus grande variété d'enjeux éthiques [24], sans nécessairement négliger la certification finale du SAM [54]. L'éthique processuelle ouvre ainsi la possibilité d'une réflexion sur le choix des procédures éthiques à engager.
Ce retour au présent et aux pratiques telles qu'elles se font n'est en rien original dans la littérature de l'éthique des techniques [20]. Il nous semble toutefois intéressant de spécifier quelques pistes pour imaginer l'intégration de cette éthique au *design* des SAM. Ce projet de recherche est particulièrement large, nous souhaitons donc ici présenter dans un premier temps les principaux lieux qui nous paraissent propices à une réflexion éthique processuelle dans le *design* et la conception de SAM.

3.1.2 Comment intégrer l'éthique processuelle ?

L'éthique processuelle ne se substitue pas à un cadre éthique préexistant. Elle constitue plutôt une manière de décliner une posture éthique particulière [25]. Il s'agit donc pour nous d'interroger avant tout la façon dont l'éthique se fait plutôt que les principes qu'elle défend. En cela, elle reste compatible avec l'approche dominante d'éthique « par principes ». Cette éthique par principes vient souligner le besoin que les résultats des SAM respectent certaines valeurs fondamentales. Ces principes regroupent tout aussi bien des caractéristiques techniques intrinsèques aux algorithmes que des valeurs liées aux usages qui en sont faits. Parmi ceux qui sont saillants dans le champ de l'IA, on compte no-

21. Par ailleurs, Nurock *et al.* ont également défendu une éthique processuelle à intégrer au *design* [42]. Contrairement à notre perspective, leur compréhension de l'éthique processuelle est marquée par celle du *care*.

22. Il s'agit pour elles d'insister sur le besoin de réflexivité dans les pratiques de l'AM, notamment en portant un regard critique sur les limites éthiques et pratiques des données employées et des modèles construits. Elles soulignent néanmoins la nécessité de construire une méthodologie propre à chaque situation.

23. Il est à souligner que ces catégories d'une part procédurale et d'autre part processuelle recouvrent en elles-mêmes une diversité de pratiques que nous ne déclinerons pas ici en détail.

24. À ce propos nous renvoyons notamment à une proposition méthodologique destinée à la méthode agile [56].

25. Ainsi que le suggère explicitement B. Mittelstadt lors qu'il parle d'encourager la vision de « l'éthique comme un processus et non pas comme un solutionnisme technologique » (« *Pursue ethics as a process, not technological solutionism* » (notre traduction)) [37, p. 10].

tamment la transparence, la non-malfaisance, l'équité et la justice, la responsabilité et le respect de la vie privée [26]. Des propositions émergent pour montrer en quoi ces principes éthiques sont mis en jeu dès le développement des SAM, et non plus seulement dans leurs usages. Morley *et al.* recensent ainsi une centaine d'initiatives pour intégrer l'éthique au sein de la phase de conception [40]. Ils soulèvent cependant le manque d'applicabilité directe des propositions issues de la littérature. Une difficulté réside dans la transcription de principes au sein du *design* des SAM. Des risques surgissent notamment d'une instrumentalisation de ces principes à des fins mercatiques et commerciales par des organisations qui voudraient paraître plus éthiques qu'elles ne le sont réellement [15].

3.2 Intrications entre technique et éthique

Nous nous proposons d'expliciter, à chaque étape, une partie des questionnements éthiques qui, bien que parfois implicites ou non-considérés, sont effectivement compris dans ce processus de conception [27]. Il est à noter que la succession d'étapes que nous allons proposer est avant tout schématique et peut être rompue au moins de deux manières. D'une part, les frontières qui les délimitent sont poreuses, c'est-à-dire que deux opérations présentées distinctement peuvent en réalité être réalisées simultanément, soit en parallèle, soit au sein d'une activité de travail. Lorsqu'une équipe travaille sur un projet utilisant un algorithme d'AM, il n'est en effet pas rare que le codage de celui-ci débute avant même que les données ne soient récoltées. Des allers-retours entre ces étapes sont d'autre part monnaie courante : il est habituel que des résultats donnés par un modèle prototypique mettent en lumière un échantillon aberrant de la base de données (BdD), qui en sera ensuite évincé. La liste d'étapes proposée n'est donc pas à prendre comme une représentation littérale de la procédure de *design*. Il s'agit plutôt d'une esquisse simplifiée à des fins analytiques. Ce découpage en étapes d'un processus continu permet d'identifier des points de vigilance qui échapperaient à une approche procédurale de l'éthique (qui se niche souvent uniquement dans la certification finale du SAM). L'enjeu est de favoriser une posture réflexive en reconnaissant les situations qui peuvent être dommageables au point de vue éthique.

3.2.1 Formalisation du problème

Le processus de conception débute au moment où il est décidé qu'il serait judicieux de recourir à un SAM pour résoudre un problème identifié. Bien que cette décision ne soit pas en elle-même une étape du processus de conception et qu'elle n'appartienne pas aux développeurs eux-mêmes en général, nous estimons qu'elle en constitue son point d'entrée et qu'à ce titre elle mérite qu'on lui porte une attention particulière. La décision de recours à l'usage de ces systèmes se fonde parfois sur les croyances que les SAM, capables de traiter un grand nombre d'informations à la fois, pourraient résoudre plus efficacement que des humains la majorité des problèmes [7]. Il appartient dès lors aux développeurs d'interroger la pertinence de l'emploi d'un SAM pour résoudre le problème réel identifié. S'ils sont les mieux placés pour définir ce que peut un système informatique, ils doivent pouvoir tout autant être critiques de ce que ce dernier ne permet pas d'accomplir [11].

Le problème réel est traduit en un problème mathématique que doit résoudre le SAM. Ce problème mathématique consiste, pour le dire simplement, en l'optimisation d'une fonction de coût. Les paramètres de cette fonction sont des variables, quantitatives ou qualitatives, qui correspondent à des attributs formalisés du problème. Pour l'évaluation future du modèle entraîné, des métriques de performance dudit modèle sont également définies. Pourtant, traduire un problème réel en un problème qui soit soluble par un algorithme n'est absolument pas évident. Une difficulté d'ordre éthique apparaît à ce stade, la correspondance entre la réalité du problème et la fonction de coût. Pour l'illustrer, un système chargé de différencier des images de chiens et de loups peut être amené à se fier à la couleur de l'arrière-plan (blanc pour les loups, à cause de la neige) plutôt qu'aux phénotypes de ces animaux [46]. Dans ce cas, le système résout bien mathématiquement le problème, mais cela ne correspond pas aux attentes qui pouvaient être investies en lui. Cela montre l'existence possible d'une divergence entre le problème formulé symboliquement — distinguer des animaux — et computationnellement — distinguer des matrices de pixels représentant pour l'humain des photos d'animaux. Cette problèmatique est d'autant plus prégnante lorsque l'on touche à des qualités humaines, qu'elles soient physiologiques ou morales. En ce sens, les variables qui servent de critères aux algorithmes de recrutement supposés trouver le meilleur salarié à attribuer à une entreprise donnée paraissent difficiles à définir précisément, et d'autant plus à quantifier [43]. C'est ce que B. Chin-Yee et R. Upshur appellent le problème phénoménologique de l'IA, à savoir l'incapacité à rendre compte de tout un pan de l'expérience humaine de manière quantitative [6]. Le choix, la définition et la mesure de ces variables deviennent dès lors des enjeux éthiques à part entière.

3.2.2 Préparation des données

Une grande partie du travail de conception réside dans la préparation de l'ensemble des données qui serviront à entraîner puis valider le modèle [28]. Cette étape détermine ce qu'apprend le modèle et les résultats qu'il obtient.

Constitution de la base de données. Dans certains cas, la BdD existe déjà en accès libre et est récupérée, sa constitution ne demande ainsi pas de captation des données. Celle-ci requiert une méthode rigoureuse pour assurer que les données soient collectées dans des conditions similaires.

26. Ce sont en tout cas ceux qui sont les plus mis en avant dans la multitude de guides éthiques proposés par des institutions, qu'elles soient gouvernementales, supra-nationales, académiques ou privées [25].

27. Les phases schématiques de développement d'un SAM proposées ici sont similaires au modèle classique CRISP-DM [4].

28. Il est à noter que la prise en compte des conditions de mise en existence de ces bases de données peut constituer une préoccupation éthique à part entière. Les activités professionnelles de micro-travail, relativement précaires, ne sont pas encore strictement encadrées et consistent en la réalisation de tâches d'annotation ou de vérification fastidieuses, pas toujours claires et qui peuvent être refusées sans motif manifeste par les clients [34].

Quelle que soit la méthode de collecte il reste important d'éviter la perte de sens qui pourrait advenir en isolant les données de leur contexte de production [32].
La connaissance d'un phénomène dépend des données choisies pour le représenter et le mesurer [43]. Une BdD satisfaisante doit être représentative du phénomène qu'elle prétend décrire, ce qui n'est pas un objectif trivial à atteindre lorsque ce phénomène n'est pas entièrement connu. Et même lorsque l'on connaît les propriétés statistiques des variables représentant un phénomène, constituer une base qui les respecte reste une tâche complexe [29]. Un manque de représentativité de la BdD introduit dès lors un risque de biais qui peuvent se manifester de manière ostensible lors de la mise en pratique du modèle dans l'espace social.

Annotation de la base de données. Dans le cas de l'apprentissage supervisé, il y a une volonté de réemployer des catégorisations théoriques ou statistiques pré-existantes. La BdD est annotée en fonction de ces catégories, le plus souvent manuellement — il y a donc une dépendance aux capacités et à la subjectivité de l'annotateur. Le soin apporté à l'annotation permet de tenter d'éviter la génération de biais, ou devrait du moins rendre compte des limites de la catégorisation effectuée. Une part importante d'arbitraire reste pourtant incluse dans cette étape de classification, sans qu'elle ne soit explicitement décrite en général [30].
La finesse de la distribution des classes attribuées aux échantillons de la BdD peut avoir des conséquences directes sur les résultats d'un SAM. Une illustration est donnée par la mort d'E. Herzberg, causée en 2018 par une reconnaissance d'obstacle incorrecte de la part d'une voiture Uber quasi-autonome [31].
L'annotation de BdD, tout comme sa constitution en général, n'est ainsi pas une activité purement descriptive (épistémique), elle a un pouvoir normatif (éthique) sur l'emploi fait des données. Éthique et épistémologie se rejoignent dans l'exigence d'une correspondance adéquate entre la BdD, son annotation et la réalité. L'évaluation de cette correspondance repose sur l'explicitation de l'origine et de la nature des données et contribue à éclairer au mieux possible les décisions éthiques.

Nettoyage et standardisation de la base de données. Cette procédure consiste à opérer des modifications de la BdD pour s'assurer de la qualité et de la correction de celle-ci. Bien que des systèmes de correction automatisée existent, l'implication d'un humain dans cette tâche est quasiment toujours nécessaire [22]. Grâce à la visualisation et à l'analyse superficielle des données [32], il est possible de trouver des échantillons qui sont considérés comme étant inadéquats. Il peut s'agir d'erreurs de mesure, de retranscription ou de disparités liées à l'utilisation de données provenant de sources différentes. Par ailleurs, les données peuvent subir un pré-traitement pour rendre possible leur analyse statistique par le système [33]. Cette calibration est également nécessaire lorsque des données sont issues de sources différentes avec leurs normes spécifiques [34].
Le nettoyage des données implique de classifier les données entre celles qui seraient correctes et incorrectes, puis d'effectuer des modifications ou des suppressions de données inadéquates. Ce sont les perceptions seules des *designers* sur ce qui correspond ou non au réel qui les amènent à opérer des transformations manuelles de la BdD. Ce qui est préjudiciable ici est d'écarter une donnée sous prétexte qu'elle apparaît anormale alors qu'aucune erreur particulière n'a été commise dans sa collection et qu'elle pourrait légitimement prétendre à être conservée. On observe ainsi à cette étape un risque à faire dériver la normativité de la normalité [5]. Dans le cas d'une BdD de petite taille, il n'est pas inhabituel de devoir effectuer des modifications sur certaines données afin de pouvoir les y inclure [35], ce qui introduit une part importante d'arbitraire et fausse la correspondance de cette donnée au réel. Un besoin d'expertise est ainsi nécessaire pour nettoyer avec soin la BdD. Cette expertise ne saurait être attribuée à un concepteur de SAM dont les connaissances sur le phénomène modélisé peuvent être limitées. Des spécialistes du phénomène en question pourraient cependant être associés au projet pour le rendre proprement transdisciplinaire dans une perspective de *design* participatif.

3.2.3 Entraînement du modèle

Durant la phase d'entraînement, les paramètres du modèle sont optimisés par des successions d'opérations mettant en jeu une fraction des données d'entraînement. Ce modèle est mis à l'épreuve sur la base de performances définies par des métriques choisies à l'avance. Un point d'attention concerne le choix de la métrique de performance. La manière dont est mesurée la performance d'un système doit être en accord avec l'usage qui sera fait de celui-ci. Ce problème est très frappant dans le cadre du diagnostic d'une maladie grave par imagerie médicale [28]. Il n'existe pas dans ce cas de critère objectif pour arbitrer entre une minimisation des faux positifs ou des faux négatifs. Le choix doit s'effectuer sur la base du contexte d'usage du SAM (clinique ou recherche ?) et de l'éthique associée (déontologique ou utilitariste ?). Il y a donc un enjeu d'alignement entre les hypothèses sous-jacentes à l'entraînement et le cadre éthique dans lequel le SAM sera employé. La réflexion qui entoure le choix d'une métrique de performance

29. Le commentaire effectué par Deschamps *et al.* [10] reproche par exemple à Asselborn *et al.* [1] d'avoir construit un dispositif de diagnostic automatisé de la dysgraphie biaisé sur le plan de la représentativité. Dans ce cas de figure, les enfants dysgraphiques de la BdD sont considérés comme étant dysgraphiques à un niveau trop élevé par rapport à l'ensemble des enfants dysgraphiques.

30. Des initiatives sont proposées afin de pallier ce problème [23].

31. La voiture était pilotée par un SAM. Parmi les causes de l'accident, l'enquête a révélé l'incapacité du programme de reconnaissance d'obstacles à attribuer une catégorie adéquate à M^me^ Herzberg, retardant ainsi le déclenchement de la procédure de freinage d'urgence [21]. La classe "piéton à côté de son vélo" n'était pas prise en compte jusqu'ici dans les différentes annotations et l'algorithme n'avait donc pas été entraîné avec des exemples qui représentaient ce cas de figure précisément.

32. Par exemple, par l'examen des distributions d'une certaine variable.

33. Typiquement, les variables peuvent être centrées et réduites pour éviter de devoir comparer des variations sur des plages d'étendues trop différentes.

34. C'est notamment le cas pour des mesures temps-réel effectuées avec des matériels dont la fréquence d'échantillonnage diffère.

35. Ramener arbitrairement à zéro, ou à la moyenne des valeurs.

constitue donc également un véritable enjeu éthique. Pour le résoudre, une démarche de *design* participatif incluant les futurs usagers paraît tout à fait adaptée. Ceux-ci peuvent par eux-mêmes donner conscience aux *designers* des besoins liés à la mise en application concrète du dispositif. Dans une démarche d'éthique processuelle, la présence de différentes parties prenantes permet d'interroger des enjeux éthiques inhérents au projet dont les développeurs n'ont pas nécessairement conscience [36].

3.2.4 Test du modèle

Des données non utilisées dans la phase d'entraînement servent à tester le modèle pour vérifier s'il est performant sur de nouvelles données, évaluant ainsi son pouvoir de généralisation. L'étape de test du modèle sert à éviter le surapprentissage, c'est-à-dire que le modèle n'apprenne à résoudre le problème déterminé que sur le jeu de données qui l'a entraîné. Cependant, si l'ensemble de test provient de la même source que l'ensemble d'entraînement, il est possible qu'ils comportent tous deux des biais similaires, même s'ils sont bien disjoints [37]. Des problèmes de performances ou des biais peuvent alors émerger lors de la transplantation du système vers un autre contexte d'usage. Une approche de *metadesign* où le système serait fait pour laisser le choix de régler l'optimisation du modèle selon la demande du contexte pourrait permettre d'éviter certains de ces écueils. Il convient alors de tester l'adéquation du modèle dans toutes les configurations possibles.

3.2.5 Intégration à un dispositif sociotechnique

Le modèle fonctionnel est inclus dans un dispositif au travers duquel l'usager pourra interagir avec lui. Il est rendu accessible par la construction d'une interface logicielle et/ou matérielle. Cette dernière peut inclure des modules à même de fournir des explications à l'usager [38].

De nombreuses questions se posent ici autour de l'interaction humain-machine et de la manière dont elle permet — ou non — le respect de principes éthiques désirables. La préservation de l'autonomie de jugement et d'action des usagers en lien avec l'automatisation de la tâche est notamment un point crucial [47]. L'usager peut également être dupé par des explications qui seraient inexactes ou fallacieuses [26], et être amené à prendre des décisions erronés sur la base de preuves insuffisantes ou trompeuses [38]. Ces enjeux — parmi d'autres — renvoient notamment aux principes d'explicabilité et de transparence des algorithmes.

Résoudre ces enjeux doit pouvoir amener à une plus grande maîtrise des dispositifs basés sur l'IA par les usagers [39]. L'approche du *metadesign* peut permettre d'accomplir un tel objectif, et ainsi éviter d'autres dérives, comme le paternalisme éthique auquel un processus de *design* trop strict pourrait mener, comme le suggère L. Floridi [14].

3.2.6 Déploiement et usage

Le système est mis à disposition de ses usagers et utilisé dans un contexte particulier. Les préoccupations en matière d'éthique de conception des SAM dépendent en partie des usages qui en découlent. Bien que l'usage qui est fait d'un système peut être indépendant de la volonté de ses *designers* [40], les éventuelles dérives doivent être anticipées et prises en compte au mieux possible [41]. En particulier, il apparaît crucial de se questionner sur la contribution apportée par un SAM dans le cadre d'une prise de décision. Les résultats prédictifs algorithmiques amènent à interrompre et diriger la procédure de prise de décision humaine collective, notamment dans le cadre hospitalier [3]. Cela pose des problèmes de gouvernance et d'attribution de responsabilité. Paradoxalement, ces résultats ne constituent pas en eux-mêmes des directives qui organisent le travail [42]. La prise en compte de la manière dont ces résultats sont présentés et justifiés paraît ainsi favoriser leur bon usage [43].

4 Conclusion

Nous avons souhaité contribuer à la discussion sur l'éthique de l'IA à partir de la notion d'éthique *by design*. Après avoir reconnu qu'elle peut s'envisager sous une pluralité de formes (dessein, définition, médiation), nous avons choisi de nous inscrire dans la caractérisation du *design* comme une relation de médiation entre les *designers*, les usagers et leurs milieux. Cette relation implique de considérer les divers degrés avec lesquels les usagers sont pris en compte dans les démarches de *design* : de l'abstraction à l'encapacitation, en passant par l'implication — du centrage sur l'usager au *metadesign* en passant par le *design* participatif. Voir le *design* comme un processus nous a permis de considérer que l'éthique — pour être *by design* — doit symétriquement revêtir une approche processuelle pour considérer l'ensemble des questionnements éthiques inhérents à la conception de systèmes techniques. Elle est en cela complémentaire d'une éthique procédurale fondée sur des évaluations ponctuelles et *a posteriori* des systèmes. Dans le cas de la conception de SAM, une telle éthique processuelle se traduit par une attention continue portée sur de nombreux points de vigilance dont nous avons fourni une liste non exhaustive. La diversité et la complexité des activités de *design* d'un SAM, aussi bien que leurs effets significatifs sur l'espace social, requièrent en effet une vision de l'éthique renouvelée qui s'intègre tout au long du processus de *de-*

36. Sur un tout autre plan, notons que l'AM demande une quantité de calculs très importante et se trouve ainsi être très gourmand en énergie. Le choix du modèle, du nombre de paramètres à optimiser, ou encore le seuil de performance à atteindre influent tous sur le nombre d'opérations en jeu, et donc sur la consommation du modèle [49].

37. En outre, comme l'ensemble de test est fini, il n'est pas évident qu'il contienne l'entièreté des situations réelles d'usage ultérieures du SAM.

38. Il s'agit en général d'explications *a posteriori*, ou *post hoc* [33].

39. C'est quelque part dans la lignée de ce que suggèrent Kraemer *et al.* [28] en plaidant pour une augmentation des potentialités de paramétrage des systèmes par l'usager.

40. On l'a vu avec l'apprentissage incrémental dans le cas de Tay.

41. Nos considérations n'occultent toutefois pas la nécessité d'un cadre éthique pour les usages de l'IA au-delà de son *design*.

42. Même si certains systèmes peuvent en eux-mêmes optimiser l'organisation du travail en automatisant la direction des salariés, comme dans le domaine de la logistique [17].

43. A. Christin remarque que de nombreux salariés s'engagent dans des pratiques de résistance routinière pour contrer et minimiser l'impact des SAM sur leur travail, dans les domaines du journalisme web et de la justice [7], ce qui n'est pas toujours concluant [18].

sign. En cela, nous croyons que notre approche est à même de dépasser le seul cadre des SAM, pour s'ouvrir à celui de l'IA en général. L'approche processuelle de l'éthique *by design* que nous préconisons peut en effet s'appliquer à tout processus de *design*. Néanmoins, notre approche analytique dans le cas des SAM permet de ne pas négliger des enjeux spécifiquement liés à l'AM, comme la possibilité d'un apprentissage continu et dynamique, ou l'inexplicabilité de certains résultats. Nous suggérons qu'une initiation à ces problématiques au sein des formations en ingénierie, ainsi qu'en science des données ou encore en santé publique, ne peut être qu'encouragée. Elle passerait d'abord par une prise de conscience de la portée éthique des pratiques de *design*. Repenser l'éthique de l'IA ce n'est pas seulement un questionnement des principes moraux ou du cadre éthique qu'il s'agit de valoriser mais c'est aussi réfléchir de façon critique à la façon de les mettre en œuvre.

Remerciements

L. D. a travaillé sur ce texte au sein de la chaire « Éthique & IA » soutenue par l'institut pluridisciplinaire en IA MIAI@Grenoble Alpes (ANR-19-P3IA-0003). L. D. a reçu, *via* le projet européen StorAIge, des fonds de l'Entreprise Commune ECSEL (EC) sous le numéro d'agrément n°101007321 [44]. É. P. a reçu le soutien financier du CNRS à travers les programmes interdisciplinaires de la MITI. Nous tenons à remercier T. GUYET, T. REVERDY et T. MÉNISSIER, ainsi que les relecteurs anonymes pour leurs retours qui nous ont permis d'améliorer ce texte.

Références

[1] Thibault ASSELBORN et al. « Automated human-level diagnosis of dysgraphia using a consumer tablet ». In : *npj Digital Medicine* 1.42 (2018), p. 1-9.

[2] Kristine BÆRØE, Maarten JANSEN et Angeliki KERASIDOU. « Machine Learning in Healthcare : Exceptional Technologies Require Exceptional Ethics ». In : *The American Journal of Bioethics* 20.11 (2020), p. 48-51.

[3] Simon BAILEY et al. « Dismembering organisation ». In : *Current Sociology* 68.4 (2020), p. 546-571.

[4] Peter CHAPMAN et al. *CRISP-DM 1.0 : Step-by-step data mining guide*. SPSS, 2000.

[5] Arthur CHARPENTIER. « L'éthique de la modélisation dans un monde où la normalité n'existe plus ». In : *Risques* 112 (2017), p. 117-121.

[6] Benjamin CHIN-YEE et Ross UPSHUR. « Three Problems with Big Data and Artificial Intelligence in Medicine ». In : *Perspectives in Biology and Medicine* 62.2 (2019), p. 237-256.

[7] Angèle CHRISTIN. « Algorithms in practice ». In : *Big Data & Society* 4.2 (2017).

44. L'EC est soutenue par le programme de recherche et d'innovation Horizon 2020 de l'Union européenne et par la France, la Belgique, la République tchèque, l'Allemagne, l'Italie, la Suède, la Suisse et la Turquie.

[8] Brandt DAINOW et Philip BREY. *Ethics By Design and Ethics of Use Approaches for Artificial Intelligence*. Sous la dir. d'EUROPEAN COMMISSION DG RESEARCH AND INNOVATION. 2021.

[9] Richard DAZELEY et al. « Levels of explainable artificial intelligence for human-aligned conversational explanations ». In : *Artificial Intelligence* 299 (2021), p. 103525.

[10] Louis DESCHAMPS et al. « Methodological issues in the creation of a diagnosis tool for dysgraphia ». In : *npj Digital Medicine* 2.36 (2019), p. 1-3.

[11] M. C. ELISH et danah BOYD. « Situating methods in the magic of Big Data and AI ». In : *Communication Monographs* 85.1 (2017), p. 57-80.

[12] Flora FISCHER. « L'éthique *by design* du numérique ». In : *Sciences du Design* n°10.2 (2019), p. 61.

[13] Flora FISCHER. « Les normativités des technologies numériques : approche d'une éthique « by design » ». Thèse de doct. UTC Compiègne, 2020.

[14] Luciano FLORIDI. « Tolerant Paternalism : Pro-ethical Design as a Resolution of the Dilemma of Toleration ». In : *Science and Engineering Ethics* 22.6 (2016), p. 1669-1688.

[15] Luciano FLORIDI. « Translating Principles into Practices of Digital Ethics : Five Risks of Being Unethical ». In : *Philosophy & Technology* 32.2 (2019), p. 185-193.

[16] Vilem FLUSSER. *The Shape of Things*. Reaktion Books, 1999.

[17] David GABORIEAU. « « Le nez dans le micro ». Répercussions du travail sous commande vocale dans les entrepôts de la grande distribution alimentaire ». In : *La Nouvelle Revue du Travail* 1 (2012).

[18] Ari Brendan GALPER. « Accommodation-through-Bypassing : Overcoming Professionals' Resistance to the Implementation of Algorithmic Technology ». Sociologie. MIT, 2020.

[19] Benoît GODIN. « Making sense of innovation : from weapon to instrument to buzzword ». In : *Quaderni* 90 (2016), p. 21-40.

[20] Xavier GUCHET. « L'éthique des techniques, entre réflexivité et instrumentalisation ». In : *Revue française d'éthique appliquée* 2.2 (2016), p. 8-10.

[21] Andrew J. HAWKINS. « Serious safety lapses led to Uber's fatal self-driving crash, new documents suggest ». In : *The Verge* (2019).

[22] Joseph HELLERSTEIN. « Quantitative Data Cleaning for Large Databases ». In : *United Nations Economic Commission for Europe*. T. 25. 2008, p. 1-42.

[23] Ben HUTCHINSON et al. « Towards Accountability for Machine Learning Datasets ». In : *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 2021, p. 560-575.

[24] Sheila JASANOFF. *The Ethics of Invention : Technology and the Human Future*. W.W.NORTON, 2016.

[25] Anna JOBIN, Marcello IENCA et Effy VAYENA. « The global landscape of AI ethics guidelines ». In : *Nature Machine Intelligence* 1.9 (2019), p. 389-399.

[26] Harmanpreet KAUR et al. « Interpreting Interpretability ». In : *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 2020, p. 1-14.

[27] Charalampia (Xaroula) KERASIDOU et al. « Before and beyond trust : reliance in medical AI ». In : *Journal of Medical Ethics* (2021).

[28] Felicitas KRAEMER, Kees van OVERVELD et Martin PETERSON. « Is there an ethics of algorithms ? » In : *Ethics and Information Technology* 13.3 (2011), p. 251-260.

[29] Sylvain LAVELLE. « Politiques des artefacts. » In : *Cités* 39.3 (2009), p. 39.

[30] Rena LEDERMAN. « The Ethical Is Political ». In : *American Ethnologist* 33.4 (2006), p. 545-548.

[31] Raymond S. T. LEE. *Artificial Intelligence in Daily Life*. Springer Singapore, 2020.

[32] Sabina LEONELLI. *La recherche scientifique à l'ère des Big Data*. Sesto San Giovanni : Mimésis, 2019.

[33] Zachary C. LIPTON. « The Mythos of Model Interpretability ». In : *2016 ICML Workshop on Human Interpretability in Machine Learning*. 2016.

[34] Clément Le LUDEC et al. « Quel statut pour les petits doigts de l'intelligence artificielle ? » In : *Les Mondes du travail* (2020), p. 99-110.

[35] David LYELL et Enrico COIERA. « Automation bias and verification complexity : a systematic review ». In : *Journal of the American Medical Informatics Association* 24.2 (2016), p. 423-431.

[36] Tom M. MITCHELL. *Machine Learning*. McGraw-Hill series in computer science. McGraw-Hill, 1997.

[37] Brent MITTELSTADT. « Principles alone cannot guarantee ethical AI ». In : *Nature Machine Intelligence* 1.11 (2019), p. 501-507.

[38] Brent MITTELSTADT et al. « The ethics of algorithms : Mapping the debate ». In : *Big Data & Society* 3.2 (2016), p. 205395171667967.

[39] Stephanie L. MOORE. *Ethics by design*. HRD Press Inc, 2010.

[40] Jessica MORLEY et al. « From What to How : An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices ». In : *Science and Engineering Ethics* 26.4 (2019), p. 2141-2168.

[41] Eirini NTOUTSI et al. « Bias in data-driven artificial intelligence systems—An introductory survey ». In : *WIREs Data Mining and Knowledge Discovery* 10.3 (2020).

[42] Vanessa NUROCK, Raja CHATILA et Marie-Hélène PARIZEAU. « What Does "Ethical by Design" Mean ? » In : *Reflections on Artificial Intelligence for Humanity*. Springer International Publishing, 2021, p. 171-190.

[43] Samir PASSI et Solon BAROCAS. « Problem Formulation and Fairness ». In : *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 2019, p. 39-48.

[44] Julie PERRIN et al. « En quête d'éthique ». In : *TSANTSA – Journal of the Swiss Anthropological Association* 25 (2020), p. 225-267.

[45] Philippe PUCHERAL et al. « La Privacy by design : une fausse bonne solution aux problèmes de protection des données personnelles soulevés par l'Open data et les objets connectés ? » In : *LEGICOM* 56.1 (2016), p. 89-99.

[46] Marco Tulio RIBEIRO, Sameer SINGH et Carlos GUESTRIN. « "Why Should I Trust You ?" » In : *KDD '16 : Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, p. 1135-1144.

[47] Monika SIMMLER et Ruth FRISCHKNECHT. « A taxonomy of human-machine collaboration ». In : *AI & Society* 1 (2021), p. 239-250.

[48] Herbert SIMON. *The sciences of the artificial*. Cambridge, Mass : MIT Press, 1996.

[49] Denis TRYSTRAM, Romain COUILLET et Thierry MÉNISSIER. « Apprentissage profond et consommation énergétique : la partie immergée de l'IAceberg ». In : *The Conversation* (2021).

[50] Peter-Paul VERBEEK. « Accompanying technology : Philosophy of Technology after the Ethical Turn ». In : *Techné* 14 (2010), p. 49-55.

[51] Sandra WACHTER, Brent MITTELSTADT et Chris RUSSELL. « Why fairness cannot be automated : Bridging the gap between EU non-discrimination law and AI ». In : *Computer Law & Security Review* 41 (2021), p. 105567.

[52] Jane WAKEFIELD. « Microsoft chatbot is taught to swear on Twitter ». In : *BBC News* (2016).

[53] Langdon WINNER. « Do Artifacts Have Politics ? » In : *Daedalus* (1980), p. 121-136.

[54] Philip Matthias WINTER et al. *Trusted Artificial Intelligence : Towards Certification of Machine Learning Applications*. 2021. DOI : 10 . 48550 / ARXIV.2103.16910.

[55] Steve WOOLGAR et Geoff COOPER. « Do Artefacts Have Ambivalence ». In : *Social Studies of Science* 29.3 (1999), p. 433-449.

[56] Niina ZUBER et al. *Empowered and Embedded : Ethics and Agile Processes*. 2021.