Véhicules autonomes et biais algorithmiques : deux impasses pour la pensée éthique

Marc-Antoine Pencolé

VÉHICULES AUTONOMES ET BIAIS ALGORITHMIQUES : DEUX IMPASSES POUR LA PENSÉE ÉTHIQUE

[Version finale acceptée, avant mise en page éditeur]

MARC-ANTOINE PENCOLÉ

Doctorant en philosophie au Sophiapol
Université Paris Nanterre
marc-antoine.pencole@zaclys.net

Résumé : L'éthique, lorsqu'elle répond aux questions urgentes que posent les nouvelles technologies, le fait parfois en sapant ses propres fondements. L'analyse du cadre au sein duquel sont pensées la question des véhicules autonomes et celle des biais algorithmiques donne à voir un écrasement du raisonnement éthique sur la rationalité technique. Or, il est possible de réarmer la discussion normative de tels dispositifs en allant puiser dans certaines traditions de philosophie de la technique, qui proposent de comprendre cette dernière comme un ensemble d'institutions médiatrices de l'activité humaine et des valeurs qu'elle engage.

Mots-clefs : éthique appliquée, médiations, véhicules autonomes, biais algorithmiques

Pour citer l'article : Pencolé Marc-Antoine, « Véhicules autonomes et biais algorithmiques : deux impasses pour la pensée éthique », in *Pistes. Revue de philosophie contemporaine*, n° 1 : Thierry Ménissier (dir.), *Ethique, politique, philosophie des techniques*, Paris, Vrin, 2021, pp. 177-206.

INTRODUCTION

[178]Le développement et la diffusion de procédés algorithmiques de régulation des interactions sociales bouleversent certains attendus largement partagés quant à ce qui est juste et acceptable ou non lorsqu'il en va de décisions affectant profondément la vie des personnes impliquées. Prenons le cas des véhicules « autonomes », sans conducteur : leur mise en circulation suscite des interrogations d'ordre éthique notamment parce qu'il est possible que ces automates se trouvent parfois confrontés à des dilemmes moraux qui ne peuvent de recevoir de réponse triviale. La réflexion a finalement pris la forme d'une vaste expérimentation « éthique », la *Moral Machine*, dont l'ambition affichée est d'aider à résoudre les difficultés qui naissent à l'interface de l'éthique et de la conception d'agents machiniques « autonomes ». De même, les algorithmes d'aide à la décision interviennent potentiellement dans des domaines très sensibles : ainsi COMPAS propose d'évaluer le risque de récidive des candidats à une remise en liberté conditionnelle, et le débat s'est cristallisé autour de l'identification de biais irréductibles entre lesquels ils est nécessaire d'arbitrer, chacun engageant des valeurs opposées. Or, la voie empruntée pour articuler ces inquiétudes, celle de l'éthique dite appliquée, peut sembler souvent tout-à-fait problématique, et très en-deçà des enjeux réels, et la recherche contemporaine en philosophie de l'éthique et des techniques s'est construite, pour ce qui est de ses champs les plus actifs, sur des présupposés, une méthode et des objectifs contestables, qui amènent à penser les problèmes à travers des cadres où finalement se perd la dimension *éthique* même de la réflexion. Nous proposons d'identifier quelques-unes des limites du cadre philosophique dominant ces débats à partir du cas des véhicules automatiques et des biais algorithmiques, avant de proposer d'autres manières d'évaluer et de critiquer ces-mêmes dispositifs.

LE DILEMME DES VÉHICULES AUTONOMES

[179]La *Moral machine* est une expérience lancée en juin 2016 par le Massachusetts Institute of Technology (MIT) situé à Cambridge aux USA visant à implémenter des normes morales dans le code des véhicules automatiques en cours de développement.[1] Les voitures sans pilote, dites « autonomes », sont devenues paradigmatiques de ces dispositifs techniques élaborés qui, étant en mesure de prendre des « décisions » analogues à celle d'un agent humain, suscitent un questionnement éthique inédit, tant quant à sa forme que par son ampleur. La souplesse d'adaptation de ces machines nouvelles doit nécessairement les confronter à des situations où les différentes conduites possibles engagent des valeurs ; alors les personnes en charge de

1 BONNEFON, J.-F., SHARIFF, A., RAHWAN, I., « The social dilemma of autonomous vehicles », *Science*, vol. 352, n°6293, American Association for the Advancement of Science, 24 juin 2016, p. 1573-1576.

concevoir l'algorithme qui règle leur comportement doivent faire face à des questions non triviales. Par exemple, imaginons une voiture sans pilote dont les freins ont lâché, qui se dirige à vive allure vers une intersection : dans un cas dramatique où elle ne trouve aucune trajectoire permettant de sauver tout le monde, doit-elle privilégier la vie de ses passagers ou des piétons déjà engagés sur la voie ? C'est pour offrir une réponse à ces questions de conception éthique de l'algorithme que la plateforme *Moral machine* a été mise en place : elle consiste à présenter le dilemme et ses innombrables variantes à des volontaires du monde entier pour recueillir et agréger leurs opinions en une moyenne dont on espère qu'elle représente objectivement la solution la plus morale.

A l'heure de l'analyse des résultats[2], 40 millions de réponses avaient été recueillies. Les paramètres du dilemme étaient les suivants : trajectoire initiale de la voiture, nombre des victimes, âge, genre, en attente d'un enfant ou non, apparence physique (athlétique, enrobée ou aucune des deux), statut de passager ou de piéton, statut social (médecin, cadre, sans-domicile, criminel, ou aucun de ceux-là), traversant la voie au feu rouge ou au feu vert. Ajoutons que parfois les victimes proposées étaient des chiens ou des chats. [180]Ces paramètres ont été corrélés aux données démographiques des personnes interrogées : situation géographique, âge, genre, niveau d'éducation, revenu, tendance politique (de conservatrice à progressiste) et engagement religieux (de très croyante à athée). L'analyse statistique révèle deux séries de constantes dans les déclarations. D'abord, en moyenne pour l'ensemble des participants, présenter un certain embonpoint diminue sensiblement les chances de se voir sauvé par un algorithme aligné sur ces préférences, par rapport à une personne mince. De même, les cadres l'emportent sur les sans-domiciles, les piétons qui respectent la signalisation sont préférés à ceux qui traversent au feu rouge, les femmes s'en tirent mieux que les hommes, et enfin de manière très significative, les jeunes sont favorisés par rapport aux personnes âgées. Dans un second temps, on observe des variations dans cette échelle de valeur qui sont corrélées à l'aire culturelle (par exemple la jeunesse accroît beaucoup plus les chances d'être sauvé en Amérique du Sud qu'en Asie de l'Est), au genre des répondants (les femmes sont plus enclines à choisir un algorithme qui privilégie les femmes), ou encore à divers facteurs structurels comme l'indice GINI de mesure des inégalités du pays des répondants (les pays très inégalitaires accordent plus de valeur à la survie des cadres).

Quelle est la pertinence de cette vaste étude pour l'éthique des techniques ? Les auteurs soulignent quelques limites quant à la représentativité des répondants (les hommes jeunes, éduqués et plutôt aisés étant surreprésentés dans l'échantillon), et se montrent plutôt prudents dans leurs conclusions, tout en défendant la nécessité de se rapprocher autant que possible de ces normes dans la mise en œuvre des véhicules automatiques. On pourrait naturellement rejeter d'emblée ce type de travaux en insistant sur l'abjection qui réside dans le fait de hiérarchiser la valeur de la vie des personnes sur des critères corporels ou socio-économiques. On pourrait également pointer plusieurs problèmes sérieux d'accès à la connaissance des normes éthiques qui sont ici simplement évacués. D'abord la confusion implicite entre la norme morale rationnelle et la norme statistique simplement descriptive, puisque cette expérience ne fait qu'établir une moyenne au sein d'une diversité d'opinions. Ensuite, l'équivalence présupposée entre l'opinion immédiate et la raison en matière morale. Enfin, la minimisation des effets de cadrage sur le contenu des déclarations : les différentes manières de formuler un [181]même cas déterminent sensiblement les réponses, de même qu'aborder le dilemme dans une langue étrangère[3] ou dans le confort d'un bureau en navigant sur la plateforme, plutôt que de le vivre en première personne conduit évidemment à neutraliser des affects dont on peut pourtant imaginer qu'ils déterminent l'accès au contenu de la norme.

Mais surtout, il nous faut mettre en lumière une limite intrinsèque à la démarche : les philosophes que l'abstraction du dilemme du tramway et de ses variations actuelles rassure se représentent sans doute assez mal les difficultés qui surgissent lors de l'implémentation réelle d'un principe moral dans l'algorithme d'un véhicule automatique. Outre l'hétérogénéité des situations de jugement déjà mentionnée, c'est précisément l'abstraction des dilemmes tels qu'ils sont posés sur le papier – et sans laquelle il serait si difficile de leur apporter une réponse univoque – qui rendra caduque toute tentative de les coder tels quels. En effet, les dilemmes sont formulés de manière à présenter des possibilités d'action entièrement déterminées, ce qui ne pourra jamais être le cas d'un véhicule en circulation.[4] Mettre en œuvre un principe donné à l'occasion d'un accident réel suppose que la machine dispose d'une quantité infinie d'informations : en sus de sa propre vitesse, de la trajectoire apparente des autres véhicules et des piétons et quelques autres paramètres dont on conçoit qu'ils puissent être saisis et calculés par des capteurs perfectionnés, la machine doit également anticiper les choix des autres agents présents et estimer toute une foule de paramètres absolument déterminants – le camion qui arrive en face va-t-il tourner au dernier moment ? freiner ? Se renverser ? Les piétons vont-ils accélérer leur traversée ou revenir en arrière ? Les cahots sur la chaussée ne risquent-t-ils pas de fausser une des trajectoires calculées ? En somme, aussi loin

2 AWAD, E., *et alii*, « The Moral Machine experiment », *Nature*, vol. 563, n°7729, Nature Publishing Group, novembre 2018, p. 59-64.

3 COSTA, A., *et alii*, « Your Morals Depend on Language », *PLOS ONE*, vol. 9, n°4, Public Library of Science, 2014, accessible à l'URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0094842, consulté le 16/05/2020 .

4 NYHOLM, S.,SMIDS, J., « The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem? », *Ethical Theory and Moral Practice*, vol. 19, n° 5, 2016, p. 1275-1289.

que nous projetons nos rêves de perception machinique, nous trouvons de l'incertain, des paramètres impossibles à anticiper, ce qui a pour conséquence de faire du calcul réellement effectué l'évaluation d'une combinaison de risques et non [182]une évaluation proprement morale. Dès lors que nous tentons de prolonger le raisonnement éthique conçu sous forme de dilemmes stylisés dans l'implémentation technique réelle des principes dégagés, c'est-à-dire lorsque nous prenons en pensée la place des ingénieurs, nous sommes contraints de faire un certain nombre de choix arbitraires, et de résoudre l'indétermination des situations réelles par des approximations qui sont en dernière instance tout-à-fait subjectives et contingentes.[5] Ce qui apparaissait comme une tâche urgente et nécessaire semble désormais beaucoup plus accessoire, et nous pouvons tout-au-moins mettre en doute la pertinence d'un tel programme de recherche.

L'IMPOSSIBLE NEUTRALITÉ DES ALGORITHMES

Penchons-nous sur un autre cas exemplaire : celui des biais algorithmiques. Le perfectionnement des algorithmes, soutenu par le développement des technologies de captation et de communication de l'information et la puissance des calculateurs modernes, a conduit à la conception de systèmes complexes chargés d'identifier ou de trier les personnes. Il peut s'agir d'assister le travail de la police pour repérer dans une foule des individus suspects[6] – dans un aéroport, une gare, comme aussi dans un carnaval –, de faciliter l'attribution des lits d'hôpitaux aux patients par les médecins[7], de présélectionner des candidats à un entretien d'embauche[8], de signaler des enfants potentiellement victimes de mauvais traitements[9], etc. Il est apparu [183]très vite que les algorithmes employés présentaient un certain nombre de biais, c'est-à-dire que les taux de réussite ou d'échec de l'identification n'étaient pas égaux selon le genre, la « race »[10], la classe sociale – pour diverses raisons, un algorithme peut rendre ces appartenances sociales déterminantes dans ses opérations alors même qu'il n'a pas été programmé pour en tenir compte. De tels dispositifs techniques ayant vocation à conditionner l'accès à diverses ressources, ce type de biais de sélection aboutit donc à tout une variété de discriminations.

Les journalistes d'investigation de ProPublica ont abondamment documenté la méthode algorithmique COMPAS[11], laquelle propose une aide à la décision des juges quant à la remise en liberté conditionnelle des prévenus. La prédiction algorithmique des risques de récidive s'est vite installée au cœur du débat public et scientifique sur les biais algorithmiques en raison du caractère particulièrement sensible de son objet. Il est apparu que l'usage de ce dispositif prédictif désavantageait largement les personnes noires, les Blancs bénéficiant au contraire de scores de risques tendanciellement plus bas et donc de remises en liberté plus fréquentes. L'entreprise Northpointe, qui à l'époque le détenait, s'est défendue d'avoir tenu compte de la « race » dans son code ; et pourtant, à l'examen, le mécanisme d'entraînement de l'algorithme a reconstitué seul des catégories raciales à partir des données socio-démographiques et judiciaires utilisées pour déterminer le poids relatif des différents paramètres dans la fréquence des récidives. Il faut bien comprendre que les ingénieurs ont sans doute développé en toute bonne foi, sans aucune volonté de discrimination, un programme parfaitement discriminatoire, car même si les traits raciaux ne sont à aucun moment inclus dans le code de l'algorithme, la « race » est plus ou moins fortement corrélée à d'autres paramètres (le plus évident aux États-Unis étant le lieu de résidence), et l'entraînement automatique [184]d'un algorithme consiste précisément à établir les corrélations les plus fortes : ainsi, il est tout à fait possible de discriminer un groupe racial en restant aveugle aux traits raciaux mais en classant les individus par le biais de « proxys », de paramètres qui expriment indirectement la « race ». La pratique est ancienne aux États-Unis, où dès la première moitié du XX^{e} siècle les banques ont commencé à régler leurs décisions de prêts sur le lieu de résidence des demandeurs, ce qui revenait *de facto* à exclure les Noirs ; le *redlining*, comme on l'a appelé, en est venu à désigner toutes les stratégies indirectes de discrimination. Aujourd'hui, le déploiement de médiations algorithmiques a ceci de nouveau qu'il semble reproduire et dissimuler des formes de *redlining* en dépit des bonnes intentions de leurs concepteurs.

---

5 BONNEMAINS, V.,TESSIER, C., SAUREL, C., « Machines autonomes « éthiques » : questions techniques et éthiques », *Revue française d'éthique appliquée*, N° 5, n° 1, ERES, 2018, p. 34-46.

6 DODD, V., « Met police to use facial recognition software at Notting Hill carnival », *the Guardian*, 5 août 2017.

7 BILLINGS, J., DIXON, J., MIJANOVICH T., WENNBERG, D., « Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients », *British Medical Journal*, vol. 333, n° 7563, 2006, p. 327.

8 DASTIN, J., « Amazon scraps secret AI recruiting tool that showed bias against women », *Reuters*, 10 octobre 2018, accessible à l'URL https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G, consulté le 16/05/2020.

9 CHOULDECHOVA, A., BENAVIDES-PRADO, D., FIALKO, O., VAITHIANATHAN, R., « A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions in Allegheny County. Fairness is a Process Property ! », in Conference on Fairness, Accountability and Transparency, 2018, p. 134-148.

10 L'usage universitaire du terme « race » est encore confus dans le contexte français : nous choisissons de le mettre entre guillemets pour signaler qu'il recouvre un mécanisme social d'assignation et en aucun cas une essence.

11 ANGWIN, J., LARSON, J., « Machine Bias », sur *ProPublica*, 23 mai 2016, accessible à l'URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, consulté le 16/05/2020 .

Les algorithmes en situation sont donc vecteurs de normes implicites et les conséquences de leur mise en œuvre pourrait vraisemblablement faire l'objet d'une réflexion éthique. Or, les travaux récents dans le champ émergent des FAT-ML (*Fairness, Accountability and Transparency in Machine Learning*) ont permis de comprendre que ce qui se jouait dans la conception des algorithmes quant à leur normativité relevait en réalité d'un conflit irréductible entre plusieurs concepts de justice, entre plusieurs valeurs.[12] Dans le cas de COMPAS – et c'est le point d'achoppement du débat sur les bienfaits de l'outil –, deux principes de justice s'opposent : d'un côté, celui qui a pu sembler évident aux concepteurs, qui consiste à minimiser le nombre total de récidives en incarcérant le moins d'accusés possible, donc à privilégier la sécurité de la population dans son ensemble à moindre coût ; de l'autre, celui qui exige de traiter chacun sur la base de ses actes, et non en fonction de mécanismes d'assignation à une identité de groupe. Les journalistes de ProPublica ont ainsi mis en lumière le choix utilitariste de la sécurité à moindre coût réalisé aux dépens de l'équité entre sujets de droits. Le conflit normatif est irréductible en ce sens qu'il ne peut être [185]résolu par aucune solution technique ; le développement « éthique par *design* » ne consiste alors plus qu'à expliciter les normes mises en œuvre par telle ou telle méthode de traitement de l'information.

L'angle par lequel la pensée éthique s'est emparée du problème des biais algorithmiques participe là encore d'une forme de distraction et de dévoiement de l'éthique. Si dans un premier temps, l'urgence a été d'appeler à considérer les dispositifs algorithmiques comme des facteurs de justice ou d'injustice majeurs – ce qui n'a pas pris moins de deux décennies, les premiers travaux systématiques sur la question remontant aux années 1990[13] –, des voix s'élèvent désormais pour éviter que la réflexion éthique ne se perde dans des arguties techniques stériles.[14] En effet, notons d'abord qu'à l'instar du codage de la « moralité » des véhicules automatiques, il y aura toujours un fossé entre la formulation abstraite de principes de justice et leur implémentation dans un algorithme. Un algorithme, qu'il serve de médiation entre juges et accusés, policiers et badauds, assureurs et clients, et quelle que soit la méthode de construction automatique de ses catégories opératoires, n'opère jamais qu'en soumettant à divers calculs des volumes importants de données numériques discrètes : son épine dorsale est ce qu'on appelle sa *fonction de coût* – c'est-à-dire la définition mathématique de l'efficacité choisie pour cet algorithme –, et c'est elle qui oriente tout le processus d'apprentissage machinique – à savoir la détermination des classes ou la pondération des paramètres du calcul. Or, un principe normatif, à moins d'être parfaitement trivial, n'a que peu de chance d'être immédiatement transposable sous la forme d'une fonction mathématique : cette hétérogénéité fondamentale creuse un [186]espace de traduction, de négociation entre différents acteurs, où on peut constater qu'interviennent des langages situés, des impératifs pratiques, et naturellement tout une série de décisions subjectives.[15] Il serait naïf de penser que la contingence du résultat final ne soit pas constitutive d'un tel processus et qu'elle puisse être éliminée en prenant plus de temps ou en laissant plus de liberté aux concepteurs.

Ensuite, on peut nuancer la centralité du problème des biais en identifiant leur cause générale : la stratification sociale et les inégalités structurelles. Un algorithme prédictif comme COMPAS est entraîné sur un ensemble de cas réels passés, en l'occurrence sur un ensemble de décisions de remise en liberté conditionnelle ou de mise en détention provisoire, prises par des juges sur un territoire donné, donc au sein d'un système de déterminants sociaux : les juges appartiennent sans doute à une catégorie assez particulière de la population, avec un *ethos* et une compréhension du monde déterminés, et les territoires sur lesquels ils rendent justice sont très probablement traversés par diverses fractures sociales, et autres phénomènes de ségrégation économique ou « raciale ». Or, si la réalité sociale à partir de laquelle l'algorithme est paramétré est structurellement défavorable aux Noirs (que ce soit en raison de biais racistes de l'appareil judiciaire, de mécanismes de reproduction de la pauvreté ou de ségrégation urbaine), celui-ci ne pourra que tendre spontanément vers la reproduction de cet état

---

12 KLEINBERG, J., MULLAINATHAN, S., RAGHAVAN, M., « Inherent Trade-Offs in the Fair Determination of Risk Scores », *arXiv.org*, Cornell University, 2016, accessible à l'URL https://arxiv.org/abs/1609.05807, consulté le 16/05/2020 ; MORGAN, A., PASS, R.,, « Paradoxes in Fair Computer-Aided Decision Making », dans *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, Honolulu, HI, USA, Association for Computing Machinery, coll. « AIES '19 », 2019, p. 85-90.

13 L'initiative de l'Algorithmic Justice League rattachée au MIT a été en cela décisive (cf. www.ajlunited.org). On notera également le travail séminal d'Helen Nissenbaum dans ce domaine : cf. FRIEDMAN, B., NISSENBAUM, H., « Bias in computer systems », *ACM Transactions on Information Systems*, vol. 14, n° 3, 1996, p. 330-347.

14 POWLES, J., NISSENBAUM, H., « The Seductive Diversion of 'Solving' Bias in Artificial Intelligence », *Medium*, *OneZero*, 7 décembre 2018, accessible à l'URL https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53, consulté le 16/05/2020 ; PASQUALE, F., « The Second Wave of Algorithmic Accountability », *Law and Political Economy*, 25 novembre 2019, accessible à l'URL https://lpeblog.org/2019/11/25/the-second-wave-of-algorithmic-accountability/, consulté le 16/5/2020 ; ZIMMERMANN, A., DI ROSA, E., « Technology Can't Fix Algorithmic Injustice », *Boston Review*, 12 décembre 2019, accessible à l'URL http://bostonreview.net/science-nature-politics/annette-zimmermann-elena-di-rosa-hochan-kim-technology-cant-fix-algorithmic, consulté le 16/5/2020.

15 PASSI, S., BAROCAS, S., « Problem Formulation and Fairness », *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta, GA, USA, Association for Computing Machinery, coll. « FAT* '19 », 2019, p. 39-48.

de fait. La conception d'un algorithme de ce genre prend appui sur un donné qui est lui-même biaisé : à moins d'admettre un essentialisme grossier qui associe nécessairement la blancheur de peau au respect de la loi, comment expliquer qu'à peine un tiers de la population carcérale est blanche quand celle-ci représente les deux tiers de la population du pays[16], sinon par une multitude de mécanismes structurels *a priori* défavorables aux personnes de couleur ? La technique jette ici un voile sur le social et dissimule non pas seulement les valeurs dont elle est vectrice, mais également l'origine sociale des injustices qu'elle reproduit. Ainsi, concentrer l'effort [187]théorique sur les biais algorithmiques revient à recevoir comme un donné les biais sociaux qui conditionnent ceux de la technique. Il ne s'agit pas de nier les effets normatifs plus ou moins autonomes des dispositifs techniques mais de montrer le point aveugle de ce champ de recherche.

L'ÉTHIQUE APPLIQUÉE CONTRE L'ÉTHIQUE

Ces quelques remarques nous ont permis de faire apparaître une tension inhérente à la discussion éthique des nouvelles technologies (des véhicules autonomes et des biais algorithmiques, par exemple) telle qu'elle s'est constituée ces dernières années. Il semble que le projet de détermination des normes morales de fonctionnement de tels dispositifs rencontre deux obstacles en apparence insurmontables : le conflit irréductible entre les normes d'une part, et la difficulté de l'implémentation concrète de ces mêmes normes d'autre part. On peut interpréter cette paralysie de la pensée éthique aux prises avec la technologie contemporaine de deux manières : soit comme la constitution d'un nouveau champ de problèmes, émergeant au sein du cadre de ces nouvelles technologies, soit comme un épuisement de l'éthique par une rationalité étrangère – strictement technique et non plus pratique au sens large – qui la condamne à l'impuissance. C'est cette seconde lecture que nous proposons de défendre ici. Les domaines de questionnement que nous avons présentés jusqu'ici *ne relèvent en fait plus de la pensée éthique*, quoi qu'ils en aient l'apparence : cela ne veut pas dire qu'ils n'aient pas de valeur – ils sont même tout-à-fait indispensables –, mais simplement que *la question de l'éthique des techniques se pose à un autre niveau*.

Essayons d'abord de comprendre en quoi le problème des voitures sans pilotes et des biais algorithmiques ne relève pas à proprement parler de l'éthique. Ce que suggère l'étude du codage concret des normes pratiques dans le dispositif technique, c'est que le résultat de l'action est en grande partie contingent et dépend de facteurs extérieurs au dispositif lui-même. Et même, dans le cas des véhicules automatiques, la moralité des principes se dilue dans l'implémentation jusqu'à se perdre entièrement. Ajoutons à cela que dans le cadre imposé aux concepteurs, quand bien même on [188]parviendrait à traduire authentiquement des normes pratiques générales en lignes de codes, le conflit des valeurs engagées semble irréductible. Finalement, quelle que soit la conception choisie, quelles que soient les valeurs adoptées, le résultat apparaîtra toujours largement immoral. C'est ce qui a conduit certains à considérer que dans certaines situations, le seul choix honorable est précisément de refuser de choisir si cela implique de hiérarchiser les vies[17], ou de s'en remettre au seul hasard pour décider quelles vies le véhicule doit épargner[18], c'est-à-dire qu'il peut être moins irrationnel d'ignorer certaines informations et d'introduire une dose d'opacité que de chercher à résoudre à tout prix notre incertitude morale.

Pour autant, les concepteurs *doivent* faire un choix informé, car si aucune solution, dans ce cadre, ne peut être moralement satisfaisante, il en est qui sont évidemment contraires à toute norme de justice. Par exemple, ne pas programmer de mécanisme d'évitement en cas d'accident pour un véhicule automatique, c'est faire le choix de sacrifier par défaut les piétons qui traversent devant la voiture ; or, si on ne parviendra pas à résoudre nos dilemmes éthiques, il est tout de même possible de trouver un arrangement moins catastrophique. Hegel, à propos de la décision judiciaire, avait déjà noté cela : « ce qui est plus important, c'est qu'une décision soit prise »[19] ; la détermination particulière du montant de l'amende infligée à tel contrevenant ne peut se faire sur aucune base rationnelle, et pourtant, il suffit qu'elle soit trop élevée ou trop faible pour que la décision bascule dans l'injustice, le pire étant qu'aucune décision ne soit prise. Il en est de même pour la programmation « éthique » des algorithmes : aucune option unique ne pourra être élue à l'exclusion des autres sur une base rationnelle, mais le plus irrationnel sera de n'en choisir aucune ; c'est là que se trouve donc malgré tout l'intérêt [189]des recherches sur les biais algorithmiques et les véhicules autonomes, une réflexion doit être menée mais en ayant à l'esprit qu'elle n'appartient plus au domaine de l'éthique à proprement parler. De même que l'éthique n'a rien à dire quant au montant exact de l'amende dès lors qu'elle est raisonnablement proportionnée, elle ne

16 BRONSON, J., CARSON, A.E., *Prisoners in 2017*, U.S. Department of Justice, Bureau of Justice Statistics, 2017.

17 DI FABIO, U., *et al*., *Automatisiertes und Vernetztes Fahren*, Bundesminister für Verkehr und digitale Infrastruktur, 2017. On se rapportera notamment à la proposition n°9 de la commission : « Bei unausweichlichen Unfallsituationen ist jede Qualifizierung nach persönlichen Merkmalen (Alter, Geschlecht, körperliche oder geistige Konstitution) strikt untersagt », p. 11.

18 GRINBAUM, A., « Contre la transparence : la valeur du hasard pour une machine apprenante », *Revue française d'éthique appliquée*, N° 5, n°1, ERES, 2018, p. 47-53.

19 HEGEL, G. W. F., *Encyclopédie des sciences philosophiques en abrégé*, trad. B. Bourgeois, Paris, Librairie philosophique J. Vrin, 2012, p. 541 ; et *Principes de la philosophie du droit*, trad. J.-F. Kervégan, Paris, P.U.F., 2003, p. 303.

peut se prononcer quant au détail des choix d'implémentations de normes dans un algorithme, celui-ci relevant toujours ultimement d'un arrangement technique.

Si ce n'est dans la résolution des dilemmes pratiques, à quel niveau la pensée éthique peut-elle être requise par l'irruption des nouvelles technologies ? L'investissement de ces modes de questionnement découle en fait d'une conception très particulière – et surtout très insuffisante – de la philosophie éthique, et c'est en dégageant les présupposés généraux d'une telle démarche que nous pouvons dessiner plus clairement les contours du travail théorique appelé par les évolutions techniques contemporaines. L'urgence n'est pas tant de résoudre des dilemmes abstraits que de comprendre ce que la technique fait aux normes : en tant que dispositifs de médiation de l'activité sociale, des algorithmes du genre de ceux que nous avons présentés jusqu'ici ont, si leur usage se répand, un effet sur le mode d'être des normes, sur notre rapport collectif à celles-ci. La question éthique centrale est bien celle du *devenir technique de la normativité* dans les sociétés contemporaines, et – non sans une certaine ironie – c'est parce que l'éthique est de plus en plus imprégnée d'une rationalité technique qui lui est profondément hétérogène qu'elle se fourvoie et perd les moyens de saisir ce problème au niveau adéquat.

Les débats qu'ont suscités la *Moral Machine* du MIT et les biais algorithmiques sont largement redevables de l'appareil conceptuel des *computer ethics* (parfois aussi « *computer and information ethics* »), élaborées dans les années 1980 aux États-Unis. Cette éthique s'est le plus souvent pensée comme une éthique *appliquée*[20], c'est-à-dire comme dédiée aux difficultés spécifiques posées par les nouvelles technologies. Selon James Moor, un de ses fondateurs, l'informatique consiste en un ensemble de procédés et d'outils qui ont [190]comme propriété essentielle d'être malléables, polyvalents et adaptables, et qui par conséquent représentent une extension inédite des capacités humaines d'agir, de saisir le réel, de l'analyser et de l'affecter, extension qui découvre à la pensée autant de nouveaux types d'interaction – d'autant plus que la souplesse des technologies informationnelles fait qu'elles sont en perpétuelle évolution.[21] Or, chacune de ces percées ouvre « un vide normatif [*a policy vacuum*] à propos de la manière dont il faudrait se servir des technologies numériques », et « une des tâches centrales de l'éthique de l'informatique consiste à déterminer ce qu'est notre devoir dans ces situations nouvelles, c'est-à-dire de formuler des normes [*policies*] pour guider notre action ».[22] L'indétermination normative est donc d'abord pensée sur le modèle du vide juridique – on peut d'ailleurs imaginer que ce sont des contentieux juridiques inédits qui ont suscité à l'époque cette prise en charge philosophique –, c'est-à-dire dans le cadre d'un ensemble de principes donnés et codifiés qu'il s'agirait d'interpréter ou de compléter de manière à qualifier des situations nouvelles. Il s'agit bien d'un problème d'*application*, mais il n'a rien de trivial, selon Moor, puisque « là où il y a un vide normatif, il y a aussi souvent un vide conceptuel. Bien qu'un problème d'éthique de l'informatique puisse sembler clair à première vue, en y réfléchissant on se rend compte qu'il est pris dans une véritable pagaille conceptuelle. Ce dont on a besoin alors, c'est d'une analyse qui offre un cadre conceptuel cohérent à l'intérieur duquel on puisse formuler des normes d'action ».[23] L'embarras des juristes, en effet, tenait typiquement à la manière dont le numérique imposait de redéfinir les notions de vol et de propriété (un logiciel peut-il être la propriété exclusive d'une personne ? peut-on vraiment s'introduire chez autrui en restant chez soi devant son terminal ?), etc. Le programme de recherche ainsi défini s'est ramifié jusqu'à l'élaboration de différentes méthodes de *design* articulées autour des questions normatives, ou encore de divers codes déontologiques des ingénieurs informaticiens, toujours dans l'optique de fournir des maximes pratiques directement [191]susceptibles de guider le travail de conception et l'usage des techniques.

Le problème se manifeste d'abord par l'usage du terme même d'« éthique appliquée ». On peut certes signifier par-là l'intention de plonger dans la complexité de la vie pratique concrète, mais on peut aussi se donner *a priori* un prisme d'analyse des questions normatives particulièrement étroit. Une éthique appliquée renvoie nécessairement à une éthique générale, et le risque est de subordonner le spécifique au général, selon un schéma statique, plutôt que de faire de l'un et de l'autre deux moments d'un même mouvement de pensée. Une fois les principes généraux déterminés – principes bien souvent réduits à une combinaison de déontologie et de conséquentialisme que chacun arrangera selon son goût propre –, il ne reste plus qu'à comprendre en quoi la situation nouvelle tombe sous la législation d'un principe ou d'un autre, ou dans quelle mesure l'un d'eux aurait besoin d'être amendé dans sa formulation pour s'appliquer pleinement. Il n'est pas question, à ce niveau, d'interroger les principes eux-mêmes – dans leur universalité, ceux-ci ne peuvent rien devoir à des cas particuliers émergeant au cours de l'histoire –, et il n'est évidemment pas question non plus d'interroger le cadre au sein duquel le vide normatif est constaté. Les concepteurs de la *Moral Machine* considèrent par exemple comme acquis que les véhicules automatiques sont une bonne chose (puisqu'ils causent statistiquement beaucoup moins d'accidents que des conducteurs humains), et la tâche de l'éthique est pour eux d'améliorer « l'acceptabilité » de la technologie en question auprès du public, ou encore de surmonter les « obstacles

20 BYNUM, T., « Computer and Information Ethics », dans ZALTA, E.N. (éd.), *The Stanford Encyclopedia of Philosophy*, Stanford University, 2015, accessible à l'URL https://plato.stanford.edu/entries/ethics-computer/, consulté le 16/05/2020.

21 MOOR, J. H., « What is computer ethics? », *Metaphilosophy*, vol. 16, n°4, 1985, p. 266-275.

22 *Ibidem*, p. 266 (nous traduisons).

23 *Ibid*.

psychologiques » à son adoption la plus large.[24] Restreindre le cadre de la discussion fait naturellement peser le risque de mettre la réflexion théorique au service d'un « *ethics washing* » tout-à-fait stratégique pour les industriels dont la santé économique dépend précisément de la diffusion sans heurts de leurs produits : la multiplication des tables rondes sur « l'intelligence artificielle éthique », le « *design* éthique », la « transparence algorithmique » repose sur une forte demande de la [192]part des concepteurs – et c'est légitime ! – mais aussi des acteurs politiques et industriels, lesquels offrent les moyens matériels de penser ces questions, évidemment en fixant le cadre qui correspond à leurs intérêts.[25] Que faire en réaction à ces effets de cadrage, qui sont un risque inhérent à cette conception étroite de l'éthique « appliquée », sinon réaffirmer, comme par un effet de balancier, des principes moraux universels sapés par la technique moderne, position selon laquelle les éthiques spécifiques seraient autant d'oublis de l'éthique en général, laquelle commanderait de s'opposer à la logique technicienne, sur une base humaniste par exemple ?[26]

Ensuite, et plus généralement, la partition conceptuelle sur laquelle opère cette éthique appliquée reste prisonnière de catégories et de distinctions statiques. Non seulement le caractère « appliqué » de l'éthique peut impliquer la position d'une éthique générale dont les principes sont pensés comme indépendants de l'évolution historique de la pratique, de ses moyens, de son outillage, mais son mode de raisonnement privilégié est celui des dilemmes, des situations dont les paramètres sont étroitement fixés et où la question du bon choix est posée à un sujet dont la position et les options sont déjà bien déterminées. Envisagée ainsi, l'éthique ne se donne plus les moyens de penser le devenir de la pratique en général et se limite à la résolution – impossible – de dilemmes fermés. Pourtant, Moor ne manque pas d'intuitions extrêmement justes quant aux effets du dialogue de la technique et de l'éthique : il reconnaît comme fait fondamental la révolution permanente intrinsèque aux technologies contemporaines, l'irruption incessante de nouveaux procédés, de nouveaux dispositifs, l'extension ininterrompue du domaine de l'agir [193]outillé ; plus essentiel encore, il remarque que toute la difficulté de résolution des cas nouveaux vient de ce que l'évolution de la pratique se répercute sur le plan des concepts par lesquels on la pense. Alors qu'il cerne avec lucidité la dynamique immanente à la technique et à la réflexion sur celle-ci, il ne va cependant pas jusqu'à définir la tâche de l'éthique en fonction. Il y a dans cette mosaïque de la recherche en éthique des nouvelles technologies comme une contradiction entre la démarche d'application de principes fixes à des cas déterminés et le constat qui motive la constitution de ce domaine, à savoir celui d'un bouleversement de la pratique et de ses catégories.

LA TECHNIQUE COMME INSTITUTION MÉDIATRICE DES NORMES

Or, on peut affirmer que la théorie éthique a été historiquement traversée par la conscience à la fois de ce qui fait son urgence et des impasses dans lesquelles elle risquait de se perdre, c'est-à-dire que les théoriciens classiques ont en quelque sorte évacué d'avance l'approche éthique dominante des véhicules automatiques et des biais algorithmiques comme étant vaine du point de vue de l'éthique – et précisons à nouveau que si la philosophie éthique n'a rien à en dire, cela ne signifie pas que ces questions ne sont pas cruciales pour nous. Rappelons les raisons profondes de la vanité de l'éthique des dilemmes pour mieux cerner en creux l'autre dialogue possible entre éthique et technique. La pensée éthique s'est toujours abstenue – au moins jusqu'à très récemment – de considérer que la résolution des dilemmes moraux, du genre de ceux qui se posent pour les biais algorithmiques ou les véhicules sans conducteurs, pouvait contribuer au progrès vers la vie bonne. En effet, se demander « qu'est-ce qui est juste dans telle situation déterminée ? » n'est pas du tout équivalent à se poser la question de ce qui fait qu'une vie peut être belle, juste, digne d'être vécue. D'un côté l'interrogation porte sur la valeur relative de différentes options d'action déterminées, de l'autre, elle englobe la totalité de la pratique dans sa concrétude, tout le processus de construction des sujets dans leurs rapports les uns avec les autres.

[194]Par exemple, si la philosophie pratique avait, dans l'Antiquité, l'ambition d'offrir des maximes d'action propres à guider le sujet dans l'édification de lui-même, elle était indifférente aux dilemmes moraux abstraits. Aristote est à ce propos plutôt explicite : les dilemmes marquent la limite de l'éthique, et il renvoie au juge le fardeau d'établir ce qui, dans l'action concrète, mérite l'éloge ou le blâme. Il est d'ailleurs tout-à-fait

24 BONNEFON, J.-F., SHARIFF, A., RAHWAN, I., « The social dilemma of autonomous vehicles », *op. cit.* ; ou encore, de manière beaucoup plus flagrante : SHARIFF, A., BONNEFON, J.-F., RAHWAN, I., « Psychological roadblocks to the adoption of self-driving vehicles », *Nature Human Behaviour*, vol. 1, n° 10, 2017, p. 694-696.

25 C'est cette convergence des intérêts économiques et des interrogations éthiques que dénonce par exemple T. Metzinger, chercheur à l'Université de Mainz, qui a participé au Groupe d'experts de haut niveau sur l'intelligence artificielle pour l'Union Européenne en 2019. Notons que le recrutement de Luciano Floridi par Google montre bien que ce qui fait l'objet d'un investissement stratégique, c'est l'affichage d'une célébrité en direction du grand public, sans réel souci de sa crédibilité académique. Cf. METZINGER, T., « Ethics washing made in Europe », *Der Tagesspiegel*, 8 avril 2019, accessible à l'URL https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html, consulté le 16/05/2020.

26 Si une réaction aussi abstraite est un lieu commun du débat public autour de la diffusion des nouvelles technologies, il n'est pas si facile d'enrôler Ellul à son service, comme on le fait parfois : celui-ci reconnaissait malgré tout les rapports dialectiques qu'entretenaient morale et technique. Cf. ELLUL, J., « Morale et technique », *Médianalyses. Cahiers de recherches communicationnelles*, n° 2, 1982, p. 24-29.

significatif que ceux-ci n'interviennent que dans la discussion de l'étendue de la responsabilité dans l'agir. On doit évidemment distinguer les actes volontaires des actes involontaires :

> Si l'acte non consenti est celui qui est exécuté sous la contrainte ou par ignorance, l'acte consenti sera dès lors, semble-t-il, celui dont le principe réside dans l'agent qui connaît chacune des circonstances particulières que suppose son action.[27]

L'idéal de l'action volontaire est celui d'un sujet qui comprend tous les détails de la situation et des conséquences de l'acte. Naturellement, cette définition révèle tout un continuum d'actions intermédiaires, réalisées en connaissance d'une partie seulement des « circonstances particulières », et dont l'agent n'est responsable que dans une certaine mesure, et pas absolument : Aristote reconnaît ainsi que les choix difficiles sont ceux qui précisément prennent place dans le domaine de ce qu'il appelle des « actions mixtes » [28], ni involontaires ni tout-à-fait volontaires, quand ni les circonstances ni les conséquences de l'action ne sont intégralement maîtrisées. Or, cette catégorie intermédiaire qui vient brouiller la distinction binaire du volontaire et de l'involontaire n'exprime rien moins que l'essence même de la pratique dès lors qu'on entreprend de la penser dans sa concrétude :

> […] les actions appartiennent au particulier et [...] le particulier est ce qui fait l'objet du consentement. Que faut-il, par ailleurs, préférer à quoi ? Pas facile de répondre, car il y a bien des différences dans le domaine du particulier.[29]

[195]En dehors de quelques cas limites, l'agir se meut dans l'élément du particulier[30] et doit composer avec son ancrage dans l'enchevêtrement des circonstances, toujours différent d'une situation à l'autre : les maximes générales par lesquelles on peut espérer l'orienter seront toujours impuissantes à elles seules à déterminer la décision la plus juste, précisément en raison de leur généralité. Aristote, en d'autres mots, ne cherche pas à abstraire l'action de sa situation, de son enracinement dans le particulier et affirme l'*indétermination morale constitutive de la pratique*. S'il est des cas, nombreux, où l'identification du devoir ne pose guère de difficulté, il en est d'autres, irréductibles, où le juste et l'injuste sont impossibles à déterminer *a priori* sur la base de principes univoques. Le dilemme ne fait alors que signaler la limite de la réflexion éthique à proprement parler et la difficulté de la tâche qui incombe au sujet moral d'un côté, et au juge de l'autre.

Hegel thématise plus systématiquement encore ces limites, et identifie la démarche qui permet de donner plus de corps à l'éthique en dépassant celles-ci. D'abord, l'agir moral est soumis à un principe d'indétermination du point de vue de son contenu objectif, de ce qu'il sera réellement une fois mis en œuvre : « Quelles sont les conséquences *contingentes* et les conséquences *nécessaires,* [cette question] contient de l'indéterminité »[31], puisque depuis l'abstraction d'une certaine situation d'action (celle de la conception d'un dispositif algorithmique donné par exemple), nos volontés ne peuvent que se heurter à l'objectivité dans laquelle elles doivent se réaliser comme à un matériau opaque et résistant, et l'étendue de notre maîtrise sur elle, et sur les conséquences de nos décisions, ne peut qu'être limitée. Ensuite, ignorons cette limitation structurelle en imaginant une maîtrise parfaite : nous buterions immédiatement sur une autre difficulté non moins insoluble, celle de l'indétermination du contenu subjectif de l'agir, de ce en quoi consiste exactement le devoir. En effet, le concept de « Bien », ce que vise notre devoir, est lui-même doublement indéterminé. Du point de vue des sujets et des situations particulières, il est naturellement possible que différents devoirs, différentes obligations s'opposent, sans qu'aucune ne [196]l'emporte sur les autres : ainsi qu'est-ce qui pourrait par exemple justifier que la vie d'un passager soit considérée supérieure à celle d'un piéton, ou bien que la prévention de la récidive dépasse le respect de l'autonomie individuelle, sinon les goûts et les couleurs d'une époque, d'une culture, d'un groupe social particulier ou de la moyenne de la population, soit tout le contraire de ce que l'on attend d'une fondation morale rationnelle de l'agir.[32] Et si l'on cherche à seulement appliquer des principes universels, c'est-à-dire découlant du seul concept de Bien et de l'obligation qui en découle, indépendamment des situations particulières, évidemment on ne retrouve plus qu'une forme vide qui n'offre aucune direction pratiquement.[33] Par conséquent, l'ambition de déterminer *a priori* la programmation de l'algorithme la plus juste, la plus morale, paraît vaine puisque cela supposerait d'une part une maîtrise infinie des effets du fonctionnement du dispositif technique, et d'autre part de fonder moralement un choix en tranchant arbitrairement parmi des options également valables. Hegel qualifie la défense résolue, dans les situations pratiques en question, de telle ou telle

---

27 Aristote, *Éthique à Nicomaque*, livre III, 1111 a 20, traduction d'après Richard Bodéüs.
28 *Ibidem*, 1110 a 10 : « μικταὶ » : mélangées, embrouillées.
29 *Ibid*., 1110 b 5.
30 « Τὰ καθ' ἕκαστα ».
31 HEGEL, G. W. F., *Principes de la philosophie du droit*, *op. cit*., § 118 add., p. 216.
32 *Ibidem*, §125, pp. 222-223.
33 *Ibid*., §134-135, p. 229-230.

option d'expression typique de « l'entendement abstrait » : cela signifie pour lui que les différentes théories morales qui s'affrontent, et qui accordent leurs faveurs à telle modalité de calcul des scores de récidive, ou à telle hiérarchie des vies entre piétons et passagers, jeunes et vieux, etc., peuvent être défendues les unes comme les autres avec des raisons tout aussi valides et s'opposer sans fin. Or, « ce qui est plus important, c'est qu'une décision soit prise », et il faut simplement reconnaître que cette décision n'est pas du ressort de la pensée éthique : en l'occurrence, si les divers codes déontologiques et comités d'éthique peuvent bien informer la décision en éliminant des options distinctement irrationnelles, le choix final des principes implémentés ne pourra relever que d'une négociation entre les contraintes techniques, légales, et les incertitudes inhérentes à la situation.
S'il n'est guère possible d'échapper à cette loi d'indétermination de l'agir moral, quelle tâche reste-t-il encore à la pensée éthique ? Suivre Hegel dans sa présentation des errements de la moralité [197]permet surtout d'envisager avec lui un déplacement de la focale théorique. La suite de son programme éthique le conduit à concevoir comment, par quels biais, par quelles médiations, les fins humaines en viennent à se réaliser *effectivement*, de manière stable et avec une certaine nécessité, c'est-à-dire à passer dans l'être au-delà de la simple virtualité inhérente au devoir-être. Alors il devient possible de chercher à savoir si une certaine évolution technologique participe d'une transformation du mode d'être des normes, potentiellement favorable au déploiement rationnel de la liberté ou au contraire profondément aliénante.

Cette théorie des normes, attentive notamment au devenir de la normativité lorsqu'elle est médiée, prise en charge, recueillie, ou redéfinie par des dispositions et des institutions de la vie collective, est en mesure de porter un regard sur la technique tout-à-fait différent de celui des *computer and information ethics*. On peut en trouver une résurgence dans une branche contemporaine de la philosophie de la technique, la « théorie des médiations techniques » représentée notamment par Don Ihde et Peter-Paul Verbeek, qui alors qu'elle est devenue une des positions standards dans le champ de la philosophie de la technique n'est même pas évoquée dans les synthèse classiques de l'éthique du numérique[34], ses présupposés fondamentaux étant sans doute parfaitement étrangers à celle-ci. A l'origine, la théorie des médiations s'inspire avec quelque liberté de la phénoménologie de Heidegger et de Merleau-Ponty[35]. Le projet de Ihde part du constat selon lequel la technique, comme les concepts, le langage ou les représentations, informe l'interaction avec le monde ou autrui qu'elle prend en charge : elle n'est pas neutre, indifférente et insignifiante, mais on doit au contraire penser le contenu de l'interaction, son sens, sa finalité ou encore les pôles engagés comme étant constitués par cette médiation. Par exemple, pour prendre des cas triviaux, le message transmis par téléphone, sa charge affective et son sens, ne sont pas ceux de la même parole si elle était reçue de vive-voix, parce que la dimension sensible de ce qui est communiqué est écrasée par le dispositif, ou encore parce [198]que la sécurité de la distance, loin du regard, peut ouvrir un espace de confidences que l'on n'oserait pas autrement. De même, le paysage contemplé par le marcheur n'est pas le même, ne se donne pas selon les mêmes modalités et ne constitue pas la même expérience esthétique s'il est aperçu depuis un train lancé à pleine vitesse. Verbeek s'est emparé de ces outils pour prolonger la réflexion sur le plan éthique[36] en interrogeant la médiation technique de l'autonomie, de l'agentivité, des valeurs, y compris à propos des technologies de l'information et de la communication[37]. L'émergence d'une technologie donnée peut ainsi bouleverser les valeurs engagées par l'activité qu'elle médie : par exemple, le développement de méthodes de diagnostic obstétrique par ultrasons permet de déceler très précocement les premiers signes d'irrégularités congénitales, dans les premiers temps de la grossesse, lorsque l'avortement est encore réalisable ; dès lors que cette technologie est largement diffusée, la relation que nous entretenions à la naissance comme destin bascule dans le domaine du choix, avec tout ce que cela implique moralement – en notant que désormais, ne pas chercher à savoir si le fœtus présente des traits neurologiques atypiques, c'est-à-dire remettre la naissance au destin, représente également un choix qui engage une responsabilité vis-à-vis de l'enfant à naître[38]. On a pu reprocher à Verbeek son refus systématique d'adopter une position critique vis-à-vis de la technique et de s'extraire de toute discussion politique pour prôner l'accompagnement ; quoi qu'il en soit, il faut lui reconnaître le mérite d'avoir réorienté l'éthique de la technique vers le questionnement du type de

34 La réactualisation de l'article de BYNUM (« Computer and Information Ethics », art. cit.) en 2018 apparaît à cet égard révélatrice.

35 IHDE, D., *Technics and Praxis*, Reidel, Dordrecht, 1979 ; et *Technology and the Lifeworld: from Garden to Earth*, Bloomington, Indiana University Press, 1990.

36 VERBEEK, P.-P., *What Things do: Philosophical Reflections on Technology, Agency, and Design*, trad. R. P. Crease, University Park (Pa), Pennsylvania State Univ. Press, 2005 ; et VERBEEK, P.-P., *Moralizing Technology: Understanding and Designing the Morality of Things*, Chicago-London, The University of Chicago Press, 2011.

37 Voir notamment VERBEEK, P.-P., « Subject to Technology: On Autonomic Computing and Human Agency », in HILDEBRANDT, M., ROUVROY, A., (éd.), *Law, human agency, and autonomic computing: the philosophy of law meets the philosophy of technology*, Milton Park, Abingdon, Routledge, 2011.

38 VERBEEK, P.-P., « Obstetric Ultrasound and the Technological Mediation of Morality: A Postphenomenological Analysis », *Human Studies*, vol. 31, n°1, mars 2008, p. 11-26.

rapport aux normes que nous voulons entretenir en nous comprenant, nous, personnes et collectifs, comme toujours déjà médiés et constitués par les assemblages techniques dans lesquels nous nous inscrivons.

[199]La théorie éthique des médiations n'est pourtant pas nouvelle : nous trouvons même quelques considérations de ce type en lien avec la technique chez Hegel, et encore une fois, prenons le temps d'un détour historique pour l'ancrer dans une perspective assez large pour armer une éthique à la hauteur des enjeux actuels. Dans ses manuscrits de Iéna[39], l'outil, l'instrument [*das Werkzeug*], est défini comme « le moyen-terme [*die Mitte*] existant rationnel, l'universalité existante du processus pratique », c'est-à-dire comme la médiation de l'activité pratique, du travail, qui lui donne son universalité en étant « ce en quoi le travail a sa permanence ». En effet, l'individu particulier engagé dans le travail est voué à disparaître avec le temps, tout comme la fin particulière qu'il poursuit, tout aussi éphémère : développer l'outil et les procédés techniques afférents, c'est prolonger l'existence du sujet qui travail, de ses gestes, de son savoir-faire, ainsi que celle de ses fins, dans un être qui lui est voué à être réutilisé, reproduit ou appris par d'autres. L'intermédiaire – la technique – accorde ainsi aux sujets et aux finalités particulières qu'il unit la permanence d'une « tradition », par conséquent, la technique porte le particulier à la double « universalité » du collectif, au-delà de l'individu, et de sa perpétuation dans l'histoire, au-delà d'une génération. De même, travailler en s'appropriant une technique, c'est apprendre à réfréner en soi l'immédiateté du désir pour s'aligner sur les méthodes et les règles du commun, héritées de la tradition et enseignées au sein d'un collectif, c'est donc se laisser constituer, soi, sujet, par le moyen de notre action lui-même. Il est tout-à-fait intéressant de noter qu'un peu plus loin, Hegel envisage le rôle structurellement problématique que peut jouer cette même technique dès lors qu'elle prend la forme de la machinerie moderne et de la division sociale du travail :

> Le travail devient d'autant plus absolument mort, il devient le travail d'une machine ; l'habileté de <l'ouvrier> singulier devient d'autant [200]plus infiniment bornée et la conscience de l'ouvrier d'usine est rabaissée au dernier degré d'abrutissement.[40]

La machinerie dérobe la dimension formatrice du travail à celui qui se conforme à ses rythmes, lequel voit son habileté diminuer au point de s'universaliser sur un tout autre mode, celui de la standardisation des forces de travail toutes substituables les unes aux autres ; collectivement, les machines sont porteuses d'une forme de désunion des groupes sociaux, leurs moyens de production de la richesse collective produisant une « populace » (la classe ouvrière) paupérisée et révoltée par la perte de valeur de son habileté et de son travail. Si le détail conceptuel du diagnostic hégélien exigerait une longue analyse, ce qui importe ici c'est de constater que la technique est saisie comme une médiation constitutive de l'individu et du collectif, c'est-à-dire qu'interroger la technique ce n'est pas – ou pas seulement – identifier les « vides normatifs et conceptuels » (cf. Moor) qu'elle découvre, mais c'est chercher à répondre aux questions : « qu'est-ce que telle technologie nous fait à nous, individuellement et collectivement ? que devenons-nous en nous alignant sur son régime ? que deviennent nos fins ? », et sans regimber lorsqu'il semble nécessaire de porter une critique profonde et politique.

La théorie des médiations technologiques de Hegel et ses prolongements dans la Théorie Critique s'avère beaucoup plus féroce que sa réactualisation tardive par Verbeek ; c'est aussi parce qu'elle est beaucoup plus ambitieuse, et assise sur une philosophie des normes plus riche. L'interrogation proprement normative de la technique chez Hegel (c'est-à-dire si on met de côté son caractère constitutif de la conscience et de la culture et qu'on l'appréhende plutôt dans l'optique de l'effectuation des normes) a toujours lieu dans son œuvre au côté de celle d'autres institutions : la famille et son patrimoine, le marché, et (pour ce qui concerne les *Principes*) l'institution judiciaire, au sein de ce qu'il désigne comme l'éthicité, ou la vie éthique [*die Sittlichkeit*]. Une éthique doit bien entendu examiner le [201]devoir des sujets moraux, comprendre ce qui assoit et ce qui limite leur responsabilité, mais elle resterait insatisfaisante et de bien peu de secours pour l'action si elle ne devenait pas également une pensée de l'institution. En effet, nous l'avons vu, l'agir moral est en partie constitutivement indéterminé, et pour le sujet confronté à la difficulté de choisir, l'exigence du Bien doit rester l'objet d'une inquiétude – ne plus s'en soucier serait régresser dans la soumission sans reste aux désirs contingents du sujet particulier –, mais ne pourra jamais être totalement satisfaite ; cependant, penser les normes ce n'est pas se contenter de les penser sur le mode du devoir-être, c'est aussi les penser dans leur effectuation. Or, celle-ci dépend moins de l'immédiateté de l'action individuelle que de leur sédimentation dans l'objectivité des mœurs et

39 HEGEL, G. W. F., *La première philosophie de l'Esprit*, trad. G. Planty-Bonjour, Paris, P.U.F., 1969. On peut également se rapporter à l'édition critique du texte original : *Gesammelte Werke*, Band 6: *Jenaer Systementwürfe I*, Hamburg, Felix Meiner, 1968, et en particulier au fragment n°20, « II. Potenz des Werkzeugs », p. 300, et au fragment n°22, p. 319-321 sur le machinisme.

40 Ces passages du fragment n°22, essentiels pour un lecteur de Smith et témoin des prodromes de la révolution industrielle allemande, ont été repris et développés dans ses cours sur la philosophie du droit lorsqu'il commente les §197-198 des *Principes* : par exemple, *Philosophie des Rechts* (1822-1823), dans les *Vorlesungen über Rechtsphilosophie*, Band 3, Karl-Heinz Itling (éd.), Stuttgart, Frommann-Holzboog, 1974, p. 602-6013, ou encore *Philosophie des Rechts* (1824-1825), dans les *Vorlesungen über Rechtsphilosophie*, Band 4, p. 500-503.

des institutions par lesquelles le sujet moral peut dépasser son face-à-face tragique avec l'objectivité du monde qu'il ne maîtrise pas.

Dans cette optique, la question éthique, celle de la vie bonne et digne d'être vécue, ne peut pas se poser uniquement au niveau du sujet moral déjà constitué et jeté face au monde, mais doit être abordée au niveau des *institutions* puisque ce sont elles qui constituent les sujets, leur fournissent les moyens matériels, cognitifs et motivationnels pour agir moralement, et qui informent le monde social dans lequel ils évoluent. L'activité humaine se déploie dans un univers saturé de normes, et celles-ci ne sont pas que le produit de décisions vertueuses portées en conscience et à chaque instant par des individualités héroïques, elles sont dans leur immense majorité mises en œuvre implicitement, sans trop y penser, en suivant la pente établie par la coutume, le bon sens hérité, les cultures professionnelles, les nécessité économiques, le Code civil, l'organisation de la justice, de l'administration, le plan des villes, l'architecture, et aussi, évidemment, par la multitude de procédés, de savoir-faire, de routines, de protocoles, de standards, d'outils, de machines qui constituent l'appareillage technique de l'activité.

Une éthique rigoureuse doit certes examiner les normes propres au *devoir-être,* mais aussi et surtout leur *devenir effectif*, leur objectivation stable productrice dans les institutions de la vie éthique, ou au contraire leur altération ou leur renversement ; en cela, puisque ces institutions sont l'étoffe du social au-delà de l'individu, toute éthique est toujours aussi une politique. Pour ce qui concerne notre objet, il nous faut considérer que *la technique est une institution* en ce sens, [202]qu'elle est un mode d'objectivation – et idéalement d'effectuation – de ce qui sans cette médiation n'existe que dans la virtualité du devoir : le questionnement éthique consiste alors à se demander ce que la médiation technique fait aux normes et à notre rapport à elles. Le passage dans la factualité d'un dispositif technique concret peut affecter de multiples manières les normes de la justice et de la vie bonne : le dispositif peut les prolonger et leur donner de l'efficace ou les dégrader, les corrompre, ou encore les réduire à autre chose, les traduire, les déplacer ; il peut donc par conséquent changer le sens d'une norme pour nous, altérer la communauté que nous formons à travers elle. C'est en considérant tous ces aspects du devenir technique des normes que l'on peut donner à entendre les enjeux éthiques réels des bouleversements technologiques actuels.

QUEL DEVENIR TECHNIQUE POUR LES NORMES ÉTHIQUES ?

Posons donc la question des véhicules automatiques et des biais algorithmiques à nouveaux frais : dès lors que nous comprenons ces dispositifs algorithmiques comme des médiations institutionnelles, que devenons-nous, nous qui interagissons à travers elles, et que deviennent nos valeurs et nos fins ?

En ce qui concerne les véhicules dits « autonomes », quelles que soient les valeurs qui sont finalement implémentées dans l'algorithme, celles-ci acquerront la factualité du dispositif qui leur donne corps. Coder ces normes et les insérer dans le programme des véhicules automatiques revient à leur donner la puissance, la stabilité et l'étendue correspondant au poids de ces véhicules dans la régulation des transports ; s'ils ne restent pas marginaux mais se diffusent au point que le trafic routier soit en grande partie automatisé, alors les normes choisies organiseront une part non négligeable de la vie collective, certes pas au niveau des grands principes moraux et légaux qui font l'objet d'une adhésion réfléchie, mais au niveau de ces règles de la socialité ordinaire qui représentent la majeure partie de la vie éthique. Or, que les valeurs qui sous-tendent ces règles soient données sur le mode de l'évidence et de la nécessité dans toute une sphère de la vie sociale n'a rien d'anodin et doit avoir comme conséquence première de les naturaliser, car il est d'autant [203]plus difficile de justifier la remise en cause d'un certain ordre normatif que celui-ci est largement réalisé dans un ordre social donné. Si par exemple du point de vue du dilemme abstrait il n'est pas plus irrationnel de laisser le hasard décider quelle vie privilégier (à nombre de vies sauvées équivalent) que de s'en remettre à l'opinion qui veut que la vie d'une personne sportive vaille plus que celle d'une obèse, en revanche il devient beaucoup plus problématique de s'accoutumer à considérer que sur la route, en général, la vie des obèses ou des vieux a moins de valeur, et plus généralement qu'il existe dans une partie de l'espace public une hiérarchie formelle et explicite établissant la supériorité de certaines vies sur d'autres. Alors la solution d'Alexei Grinbaum[41] en faveur de l'opacité et du hasard prend tout son sens : l'impératif de sauver le plus de vies possible sans faire de hiérarchies entre elles est sans doute plus adéquat à l'idéal d'égalité que nous pouvons vouloir préserver dans les autres sphères sociales. Imaginons comme il serait difficile de préserver en nous et de transmettre aux générations futures le sentiment de la validité du principe de la valeur égale de toutes les vies humaines, si celui-ci est *formellement* nié dans tout un pan de nos interactions sociales – évidemment, proclamer formellement l'égalité permet de masquer l'inégalité réelle et son effectuation ne se réduit pas à cela, mais il faut considérer que l'on franchirait un cap décisif en affirmant l'invalidité de l'idée selon laquelle chacun a droit à la vie au même titre que chaque autre, indépendamment de son âge, de son genre, de son poids, de sa couleur etc. Ainsi, une éthique des techniques sensible à la dimension institutionnelle d'un dispositif comme celui des véhicules automatiques ne peut certes

41 *Op. cit*. Pour aller plus loin, l'auteur développe largement l'option défendue dans *Les robots et le mal*, Paris, Desclée de Brouwer, 2019.

pas préférer absolument telle ou telle norme en vue de son implémentation, mais peut très clairement éliminer certaines possibilités qui sous cette lumière nouvelle apparaissent insupportables, au titre desquelles notamment l'ensemble des normes dérivées des résultats de la *Moral Machine*.

Pour ce qui est de la discussion des biais algorithmiques, il nous faut considérer deux risques : l'évolution de la responsabilité du juge et la tendance à la reproduction de l'état de fait. Tout d'abord, la [204]machine se présente à première vue comme un outil d'évaluation au service du juge chargé d'apprécier la nécessité de la détention conditionnelle, ou du médecin qui cherche l'allocation optimale des lits dont il dispose, ou de l'agent de police qui filtre les passants dans la rue ou les passagers d'un aéroport, etc. Cependant les propriétés intrinsèques des algorithmes complexes d'aide à la décision – ce qui peut aussi concerner des algorithmes relativement simples mais « privateurs », dont le code est propriétaire et les mécanismes illisibles – les rendent assez peu, voire pas du tout, transparents. La personne qui formellement est en charge de rendre son verdict doit soit acquérir une expertise certaine quant au fonctionnement de ce nouvel outil – à supposer que ce soit même possible pour qui n'est pas spécialiste de l'informatique – pour être capable d'évaluer l'évaluation qu'il propose, en connaissant ses tendances à sous-estimer ou exagérer tel ou tel facteur, sa manière de pondérer les différents paramètres, soit s'en remettre totalement au jugement de la machine sans pouvoir nuancer ses résultats au cas par cas. Si l'on ajoute à cela qu'un algorithme est aussi précis que les opérations mathématiques qui le composent, et qu'obéissant à une mécanique parfaitement déterministe il ne peut pas faire d'erreur de calcul (si par là on entend une déviation relativement à ce pour quoi il a été programmé), face à un juge toujours trop humain, passionné et inconstant, il en résulte que la machine revêt une autorité proportionnelle à celle dont le juge est dépossédé. Cet effet d'autorité affecte directement la nature de la responsabilité qui échoit au juge : parviendra-t-on à considérer qu'un jugement erroné rendu sur la base d'une recommandation algorithmique portera la même responsabilité qu'un jugement émis sans assistance, ou encore qu'un jugement rendu *contre* l'avis de la machine ? Il est à craindre que l'on n'osera plus contester le verdict des machines pour ne pas risquer d'endosser une responsabilité trop lourde en cas d'erreur. La préférence portée à tel ou tel biais, telle ou telle norme de justice à implémenter dans l'algorithme, n'est là encore pas aussi décisive que la question du type de responsabilité que nous voulons instituer collectivement. Cela ne signifie pas qu'il faut rejeter en bloc toute aide à la décision algorithmique, car parfois nous pourrions vouloir décharger quelqu'un de certaines responsabilités, ou alors le bénéfice pour nous dépassera de loin la transformation de la responsabilité de l'arbitre dans un secteur donné ; cependant dans le domaine [205]pénal par exemple, la médiation par la subjectivité du tiers est tellement constitutive de l'effectuation des normes de justice que l'on ne peut pas raisonnablement introduire une médiation algorithmique sans les vider de leur substance en croyant les porter à l'effectivité.

Ensuite, puisque les algorithmes par lesquels on régule certains processus sociaux sont entraînés sur des ensembles de données réelles, même si on n'ajoute pas de nouveaux biais à cet échantillonnage (en évitant d'entraîner un algorithme de reconnaissance faciale sur une population presque exclusivement composée de Blancs par exemple), le fonctionnement du dispositif ne pourra que reproduire de diverses manières l'état de fait existant, les partitions sociales déjà établies, les associations symboliques dominantes, les rapports de pouvoir et les inégalités de destin. Si par exemple la pauvreté est dans la réalité un facteur déterminant de la réussite des études, un algorithme complexe de sélection des candidats à une formation, même s'il n'a pas été programmé pour inclure des données relatives à l'origine sociale, reproduira certainement celle-ci comme catégorie *ad hoc* en la déduisant de données liées, qu'il s'agisse du lycée d'origine, de la profession des parents, etc. Imaginons qu'une université française soit autorisée à pondérer dans un algorithme le score des candidats à Parcoursup en fonction de leur lycée d'origine (ou qu'elle ait la capacité à le faire informellement), elle serait en mesure de discriminer pratiquement les élèves en fonction de leur statut économique pour s'assurer des taux de réussite élevés, et redoublerait les mécanismes de ségrégation sociale qui ont institué cet état de fait dans un premier temps. Or, l'opacité des mécanismes du traitement automatique des données, et la focalisation de l'attention sur ses effets de distorsion propres risquent surtout d'occulter les causes de l'état de fait reproduit et de détourner des efforts nécessaires à sa transformation. La technique bouleverse les relations sociales autant qu'elle les fige. Les normes de justice, de sûreté, de mérite, de santé publique, etc. que l'on espère porter à l'effectivité en leur conférant la puissance factuelle d'une machine risquent bel et bien de se perdre dans le processus : la médiation algorithmique d'une valeur peut, dans certains dispositifs, la mutiler pour ne la réaliser que dans le cadre de la reproduction active de l'état de fait existant, quitte à la nier dans ses fondements, comme dans le cas de la justice étayée par un outil comme COMPAS.

CONCLUSION

[206]Le détour historique par des pensées éthiques qui situent leur tâche au-delà de la résolution de dilemmes moraux abstraits permet de dégager des cadres d'évaluation plus propres à saisir les dispositifs techniques que l'on voit émerger actuellement, qu'il s'agisse des véhicules automatiques ou des régulations algorithmiques de divers processus sociaux, pour ce qu'ils sont vraiment, c'est-à-dire des *analogues d'institutions médiatrices de valeurs*, médiatrices au sens où elles incarnent et portent ces valeurs mais aussi au

sens où celles-ci se trouvent altérées, épaissies d'autres significations ou au contraire niées en passant en elles et en se (dé)réalisant techniquement.

Sur un autre plan, on peut également remarquer que les dispositifs techniques en question appellent immédiatement un certain mode de réflexion éthique dont on a pu voir qu'historiquement il n'a rien d'évident ; sans doute peut-on émettre l'hypothèse selon laquelle la forme de la pensée éthique n'est pas indifférente à la nature des objets desquels elle est sommée de s'emparer. Les impasses que représente la tentative d'arbitrer absolument entre différentes options de valeurs à implémenter résultent finalement selon nous de la confusion entre des problèmes techniques, qui peuvent faire l'objet d'un calcul prudentiel et d'une négociation situés, et les diverses déterminations du problème éthique de la vie bonne et digne d'être vécue : l'éthique est prise là dans un impasse non en raison de la difficulté redoutable des problèmes rencontrés, mais précisément parce qu'elle s'y perd *en tant qu'éthique*. La logique d'abstraction inhérente au fonctionnement d'un algorithme, sa réduction du donné à un ensemble fini de données discrètes, l'impératif de traduire une norme complexe en une fonction de coût numérique dépendant d'une série finie de paramètres, tout cela contribue à écraser la rationalité pratique dans toute son ampleur sur sa seule dimension technique, la plus à même de se couler dans le moule de la conception et du fonctionnement de ces technologies. L'éthique est donc de plus en plus appelée à se plonger dans le monde des médiations numériques, mais le *devenir technique de la pensée éthique*, s'il peut être l'occasion d'appréhender les formes nouvelles de domination à l'œuvre ainsi que les formidables opportunités d'effectuation normatives qui se présentent à nous, peut tout aussi bien conduire à son épuisement, à la perte de l'éthique dans la technique.