Une nouvelle perspective pour des technologies plus éthiques

Racisme, effusion de sang, spécisme, changement climatique… Vous voulez un monde plus éthique ? Le Dr Thilo Hagendorff, chercheur à l'université de Tübingen, affirme que vous devriez abandonner les schémas de pensée traçant une frontière artificielle entre « les siens » et « les autres » et développer une compassion inconditionnelle. Vous voulez une IA et des technologies plus éthiques ? Oubliez l'approche dominante basée sur des principes et adoptez celle basée sur des vertus. Vous devriez également commencer à travailler à la mise en place d'un climat de travail éthique.

Avant de commencer, pouvez-vous nous expliquer rapidement votre expertise qui couvre l'éthique de l'IA, l'éthique des médias et l'éthique des technologies ?

Ce sont tous des termes vagues, et ils se recoupent souvent. J'ai une formation en philosophie, mais je considère maintenant que je me suis éloigné de ce domaine, car je trouve que beaucoup de discours philosophiques sont quelque peu obsolètes. Je préfère mener des recherches transdisciplinaires incluant la sociologie, les *Science and Technology Studies* et l'éthique. Lorsque j'identifie des questions de recherche intéressantes, je me demande si je peux y travailler seul ou si j'ai besoin de l'expertise d'autres chercheurs. Par exemple, j'ai eu besoin d'aide pour rassembler des données empiriques et mener la récente étude que j'ai publiée sur la participation de l'industrie à la recherche sur l'apprentissage machine[1]. Cette étude analyse empiriquement les liens entre les institutions publiques et privées en matière de recherche sur l'apprentissage machine. Notre analyse compile les articles des trois principales conférences sur ce champ de l'IA depuis les cinq dernières années. J'espère que beaucoup de gens le liront afin que les arguments futurs puissent être mieux différenciés.

Face à la multitude de directives éthiques sur l'IA, comment pouvons-nous vérifier que les développements actuels sont réellement plus éthiques ?

C'est une question difficile et je n'ai certainement pas de réponse définitive. Je pense qu'il y a deux côtés de la même histoire qui peuvent être racontés : la version pessimiste et la version optimiste.

Le pessimiste affirme que les développements actuels de l'IA sont moins éthiques. En effet, il y a beaucoup d'*« ethics washing »* en ce moment. De nombreuses entreprises envoient des signaux forts et continus au public et aux législateurs pour leur faire sentir qu'elles gèrent l'éthique de l'IA, afin d'éviter des réglementations juridiquement contraignantes. Leur message est le suivant : « Nous avons un comité d'éthique, il n'y a pas besoin de réglementation, l'influence des normes non contraignantes est suffisante ».

D'autre part, l'optimiste souligne que l'abondance actuelle du travail éthique en matière d'IA a le mérite de sensibiliser les entreprises et la population à l'importance des valeurs en IA, et plus largement encore concernant les technologies.

Lequel de ces points de vue est vrai ? Peut-être les deux en même temps.

[1] Hagendorff, Thilo, and Kristof Meding. "The Big Picture: Ethical Considerations and Statistical Analysis of Industry Involvement in Machine Learning Research." *arXiv preprint arXiv:2006.04541* (2020). https://arxiv.org/pdf/2006.04541v1.pdf

Comment traduire concrètement les directives éthiques en développements techniques ?

Dans un article que j'ai publié au début de cette année[2], j'ai comparé 20 lignes directrices éthiques de l'IA pour analyser quelles valeurs sont les plus importantes, lesquelles sont sous-représentées et lesquelles sont omises.

Toutes ces lignes directrices tentent de différencier des principes très abstraits qui sont difficilement applicables. En outre, de chaque principe découlent des approches différentes. Prenons l'exemple du principe de la vie privée : une ligne directrice encouragera la « *privacy by design* », une autre une liste de critères à remplir pour garantir une vie privée minimale, etc. Cette approche déontologique dominante vise l'adhésion du praticien à ces principes et règles. Cependant, des études empiriques montrent qu'elle ne modifie pas leur pratique quotidienne. Par conséquent, je doute que cette approche désormais dominante soit prometteuse. Alors que cette dernière est fondée sur des valeurs qui sont des objectifs et des concepts mentaux immatériels, nous devrions envisager une autre approche.

Cette autre approche peut être basée sur des vertus qui sont des dispositions que chaque individu peut adopter et qui affectent ses actions, comme la disposition à prendre soin des autres. C'est facile de postuler des principes pour les technologies et de viser des objectifs socialement acceptables, mais il est assez difficile d'éduquer les praticiens et de renforcer leurs vertus, de sorte que, par exemple, leurs actions tendent à être plus enclines à la justice. Hormis quelques discussions académiques (lire Deborah Johnson sur la question de savoir si l'éthique peut être enseignée aux ingénieurs[3], et cette étude intéressante[4] sur le cadre nécessaire aux praticiens pour adapter leur routine quotidienne), la majeure partie de l'énergie est consacrée à l'approche fondée sur les principes, à l'élaboration d'ensembles de principes et au passage de ces principes à la pratique, alors que nous devrions essayer l'alternative fondée sur les vertus.

Les praticiens ayant de bonnes vertus peuvent-ils produire une IA éthique ?

Les principes seuls ne peuvent pas garantir une IA éthique. Nous avons besoin de vertus, mais elles restent des dispositions personnelles. Nous devons donc également mettre en place des climats de travail où les décisions éthiques ne sont pas sanctionnées, mais récompensées. Pour comprendre ces défis, je vous recommande cette grande métaétude de 2010[5] qui s'est appuyée sur plus de 30 ans de recherche pour déterminer les facteurs de choix non éthiques couvrant les dispositions individuelles, la question morale et l'environnement organisationnel (égoïste ou éthique).

Il est essentiel de garder à l'esprit les différents facteurs pour comprendre ce qui est en jeu, et que tout ne se résume pas aux dispositions individuelles, mais à plusieurs autres facteurs comme, par exemple, un climat de travail éthique.

---

[2] Hagendorff, Thilo. "The ethics of Ai ethics: An evaluation of guidelines." Minds and Machines (2020): 1–22.
[3] Can engineering ethics be taught? Deborah G. Johnson (2017)
https://www.nae.edu/19582/Bridge/168631/168649.aspx
[4] Morley, Jessica; Floridi, Luciano; Kinsey, Libby; Elhalal, Anat (2019): From What to How. An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices. In arXiv, pp. 1–21.
https://arxiv.org/ftp/arxiv/papers/1905/1905.06876.pdf
[5] Kish-Gephart, Jennifer J.; Harrison, David A.; Treviño, Linda Klebe (2010): Bad apples, bad cases, and bad barrels. Meta-analytic evidence about sources of unethical decisions at work. In The Journal of applied psychology 95 (1), pp. 1–31.
https://www.researchgate.net/publication/41087509_Bad_Apples_Bad_Cases_and_Bad_Barrels_Meta-Analytic_Evidence_About_Sources_of_Unethical_Decisions_at_Work

Quelles sont les bases de la réflexion sur l'éthique des non-humains ?

En ce moment, nous avons une vision totalement anthropocentrique sur ces sujets. Les gens ont un statut moral, mais pas les animaux ni les plantes. En même temps, les discours scientifiques comme le nouveau matérialisme ou la théorie de l'acteur-réseau disent que les artefacts techniques peuvent agir dans notre société, et ont de fait une sorte d'agentivité. C'est contre-intuitif. L'intuition accorderait aux animaux une agentivité, puisqu'ils sont plus proches de nous !

Le problème réside dans l'établissement d'une frontière mentale entre les « siens » et les « autres ». Les gens dessinent un cercle autour d'eux : ils incluent les humains, alors que les plantes, les animaux et les êtres techniques restent à l'extérieur. À l'heure actuelle, nous constatons des orientations très fortes vers les « siens ». Des études psychologiques permettent de mesurer l'« orientation de dominance sociale »[6] d'un individu, un trait qui évalue la tendance à penser en termes de groupes hiérarchisés. Cela génère le spécisme ainsi que des discriminations raciales et sexistes.

D'un point de vue pragmatique, les frontières sont nécessaires pour que les êtres humains apprennent et utilisent des termes lorsque nous essayons de saisir la réalité : elles sont à la base de notre compréhension du monde. De plus, les gens sont enclins à créer cette frontière mentale entre les groupes des « siens » et les « autres » : c'est un héritage de l'évolution.

Néanmoins, je soutiens que nous avons besoin d'une approche plus holistique pour divers sujets comme le comportement des machines, le changement climatique, les migrations, le racisme, l'agriculture industrielle et tant d'autres. Il est vraiment important d'avoir une compassion inconditionnelle envers les membres des groupes extérieurs, les « autres ». Bruno Latour, dans sa théorie du Parlement des choses, demandait une politique des artefacts. Même parmi les êtres humains, Jacques Derrida a dénoncé avec son approche déconstructiviste les nombreuses différenciations artificielles que nous faisons et qui sont à l'origine de beaucoup de sang versé.

Ces discussions fascinantes restent malheureusement essentiellement académiques, il y a encore beaucoup de chemin à parcourir pour atteindre les médias et le grand public. Pourtant, nous devons inclure la nature et les animaux dans notre éthique, le sort de la planète est en jeu.

Est-ce une bonne idée d'inclure également les robots ? Je ne sais pas. Les gens ont de la compassion pour les robots sociaux, des études ont montré que les mêmes zones du cerveau sont activées lorsque nous voyons des personnes et des robots sociaux blessés. Peut-être pouvons-nous exploiter ce sens de la compassion.

Propos recueillis par Lauriane Gorce, directrice scientifique de l'Institut de la technologie pour l'humain — Montréal

[6] Définition de Wikipedia traduite de l'anglais : « L'OSD est conceptualisé selon la théorie de la domination sociale comme une mesure des différences individuelles dans les niveaux de discrimination fondée sur le groupe ; c'est-à-dire qu'il s'agit d'une mesure de la préférence d'un individu pour la hiérarchie au sein de tout système social et de la domination sur les groupes de statut inférieur. Il s'agit d'une prédisposition à l'anti-égalitarisme au sein des groupes et entre eux. »