Tester, tester, tester les boîtes noires de l'IA

Dominique Desbois

Tester, tester, tester les boîtes noires de l'IA

Dominique Desbois

Dominique Desbois est ingénieur statisticien à l'Unité mixte de recherches « Économie publique » d'AgroParisTech et milite au syndicat CFDT de l'Institut national de recherche pour l'agriculture et l'environnement (Inrae). En tant qu'élu du personnel, il siège au conseil scientifique d'AgroParisTech et au conseil d'administration d'Inrae. Par ailleurs, il est membre du comité de rédaction de Terminal[1]*, revue de réflexion critique, qui analyse les impacts de l'informatisation sur la société. À ce titre, il est le délégué français au comité technique Information, communication & Society (TC9)*[2] *de la Fédération internationale pour le traitement de l'information (Ifip)*[3]*, organisme des Nations unies à vocation professionnelle.*

Peut-on définir l'IA en quelques mots ?

Je répondrai : non. Sans remonter à la Machine analytique de Charles Babbage[4] et au premier programme d'Ada Lovelace[5] dont les applications potentielles visaient l'astronomie et la navigation, il suffit d'un regard rétrospectif sur les principales étapes de son développement pour constater que l'IA est une véritable tour de Babel !

Il y a 70 ans, Alan Turing, co-déchiffreur du code Enigma et auteur du concept de machine universelle, pose dans la revue *Mind* la question de l'automatisation des

---

1 *Terminal* s'attache à une réflexion pluridisciplinaire sur les impacts culturels et sociaux des technologies de l'information et de la communication (TIC). Cette revue analyse les enjeux des TIC en termes de libertés individuelles et de développement durable. Pour l'édition de ses dossiers, la revue collabore avec des laboratoires universitaires, mais également avec des associations citoyennes et des contributeurs indépendants.

2 ifiptc9.org/

3 International Federation for Information (https://www.ifip.org).

4 Intégrant les travaux de Pascal et de Leibnitz sur les machines à calculer, Charles Babbage élabora les concepts fondamentaux sur lesquels repose encore de nos jours l'architecture des ordinateurs, dans un mémoire intitulé *On the Economy of Machinery and Manufactures* (Charles Knight, Londres, 1835) destiné au calcul et à l'impression automatiques des tables nautiques et astronomiques qui comportaient à l'époque de nombreuses erreurs.

5 Née Ada Byron, Augusta Ada King, comtesse de Lovelace, est l'auteure du premier programme destiné à être exécuté sur une machine (la *Machine analytique* de Charles Babbage) pour le calcul des nombres de Bernoulli en 1843.

capacités cognitives sous une forme volontairement provocatrice : « Une machine peut-elle penser ? »

Depuis ses balbutiements, les annonces médiatisées des prouesses technologiques attribuées à l'IA laissent à croire à une unité disciplinaire alors que, sous le même vocable, s'agrègent des problématiques très diverses et des techniques très différentes. Cependant, cette tour de Babel numérique qui s'étend des mathématiques théoriques à l'électronique industrielle, s'organise autour de tâches fondamentales qui fédèrent les différentes approches : représenter l'information, raisonner, résoudre et in fine apprendre.

Comment l'IA peut-elle impacter notre environnement socioéconomique ?

Au cœur de technologies comme la robotique ou la reconnaissance des formes, l'IA vient assister, suppléer, voire dépasser un certain nombre de facultés cognitives humaines de manière efficace à la fois au plan technique mais aussi économique, qu'il s'agisse de processus de production comme dans le secteur de l'automobile, d'assistance à des opérateurs humains comme pour le pilotage d'installations à risque, ou d'intermédiation comme dans le secteur de la distribution ou des services à la personne.

Est-ce que l'IA change nos vies ?

Bien sûr, et assez souvent à notre insu. Par exemple, dans le domaine de la santé. La technique de l'apprentissage profond permet désormais d'entraîner un réseau de neurones à détecter des mélanomes sur des photos de peau et des rétinopathies diabétiques sur des images de fonds d'œil.

Dans notre vie quotidienne, les « bots » se multiplient : pour trier le courrier postal[6], filtrer nos courriels indésirables ou analyser nos traces sur Internet, pour la gestion des préférences, des files d'attente et la prise de réservations sur les plateformes numériques, ces agents conversationnels capables de traiter le langage naturel (écrit ou parlé) peuvent désormais remplir certaines fonctions relationnelles. C'est le cas des auxiliaires de vie dans les établissements d'hébergement pour personnes âgées dépendantes[7].

---

6 Y. LeCun *et al.*, « *Backpropagation applied to handwritten zip code recognition* », *Neural Computation*, 1989, 1(4), p. 541-551.

7

L'automatisation des procédures de contrôle constitue également un des domaines d'application de l'IA parmi les plus dynamiques. En matière fiscale, l'utilisation de techniques d'apprentissage automatique permet de détecter les comportements récurrents qui seraient spécifiques à certains types de fraudes[8] (à la TVA, au blanchiment, à la fausse domiciliation et à l'optimisation illicite). De telles techniques sont désormais mises en œuvre par l'administration fiscale française. Dans le domaine financier, la quasi-totalité des opérateurs tels que les banques, les fonds de pension ou les investisseurs institutionnels utilisent désormais des logiciels de trading algorithmique qui mobilisent des techniques d'IA pour exécuter les transactions électroniques nécessaires à l'optimisation en continu de leur portefeuille boursier. La lecture automatique des plaques minéralogiques est désormais une application intégrée à la gestion du trafic routier, comme c'est le cas pour le péage urbain du Grand Londres. La reconnaissance faciale est couramment utilisée lors des contrôles de police, notamment dans les aéroports, et par certains pays dans la gestion des manifestations à partir de caméras de surveillance, fixes ou embarquées sur des drones, en se basant sur des techniques de reconnaissances de formes biométriques similaires à celles développées pour les empreintes digitales.

Dans le secteur de l'énergie, EDF propose à ses clients industriels Metroscope, une solution de diagnostic énergétique basée sur l'intégration de techniques d'IA[9].

Ces quelques exemples ne représentent que la partie émergée de l'iceberg dans la mesure où de nombreux projets restent confidentiels, car ils sont en phase de développement comme le pilotage automatique dans l'industrie ferroviaire, ou de test pour l'industrie aéronautique voire couverts par le secret industriel ou le secret-défense.

Et si l'IA se trompe ?

Cela arrive effectivement et les antécédents sont le plus souvent assimilables à des erreurs dans l'utilisation ou la conception des logiciels. Leurs conséquences sont loin

---

8 Édouard Pfimlin, « Zora, la solution robotique au service des seniors », *Le Monde*, 13 juillet 2018.

9 D. Desbois, « Drague fiscale sur les réseaux sociaux », *Terminal*, n° 125-126, 25 novembre 2019.

https://www.lesechos.fr/2018/03/edf-lance-sa-start-up-dans-lintelligence-artificielle-987764

d'être anodines : des biais peuvent être générés par la base des données sur laquelle on entraîne un algorithme (base d'apprentissage), ou induits par des corrélations insoupçonnées avec des règles de décision, des critères de sélection voire des comportements algorithmiques. En 2018, Amazon a été contraint de revoir son système de sélection automatique de curriculum vitæ après avoir constaté qu'il pénalisait systématiquement les curriculum vitæ féminins : en cause, la base d'apprentissage reproduisant un déséquilibre dans la distribution des genres, car constituée à partir des recrutements antérieurs sans que leurs biais n'aient été corrigés. De ce point de vue, la technologie d'IA mobilisée n'est pas neutre, car certaines méthodes tendent à renforcer les biais inscrits dans la base d'apprentissage.

En reconnaissance faciale, les problèmes de biais entachant les résultats fournis par les algorithmes d'IA peuvent aboutir à des discriminations ethniques ou sexistes. Une étude du MIT Media Lab a révélé en 2018 que ReKognition, l'IA de reconnaissance faciale d'Amazon, présentait des biais susceptibles de générer des discriminations : la détection de genre affichait un taux de réussite à 100 % pour les hommes à la peau claire alors que ce score tombait à 81 % pour les femmes, voire à 69 % pour les femmes à la peau sombre, sachant que le score de réussite d'un progiciel concurrent était de 98,5 % pour les femmes à la peau sombre.

La reproduction, voire l'amplification par renforcement, de biais non détectés a priori dans les bases de données d'apprentissage, fait de l'IA un levier de discrimination, d'autant que ces biais dits « cognitifs[10] » sont difficilement détectables a priori : c'est le talon d'Achille des algorithmes pointé par un rapport récent de l'Office parlementaire d'évaluation des choix scientifiques et technologiques(Opesct)[11].

Récemment en 2018, la Commission nationale informatique et libertés (Cnil) a mis en demeure le ministère français en charge de l'enseignement supérieur de réformer la

---

10 Initialement, le terme de « biais cognitif » a été introduit par Daniel Kahneman et Amos Tversky pour décrire les tendances observées dans les comportements humains apparaissant comme irrationnelles au plan économique.

11 Rapport d'information n° 464 (2016-2017), C. de Gannay et D. Gillot, Office parlementaire d'évaluation des choix scientifiques et technologiques, déposé le 15 mars 2017.

procédure d'admission post-bac à la suite d'une plainte introduite en 2016, afin que celle-ci respecte les dispositions de la loi informatique et libertés de 1978 sur le traitement informatisé des données personnelles. Entre autres motivations, la Cnil souligne que l'absence de détails sur les rouages de l'algorithme utilisé pour procéder au classement et à l'affectation du candidat bachelier contrevenait à l'article 39 de la loi de 1978 en ne lui permettant pas « de connaître et de contester la logique qui sous-tend le traitement automatisé en cas de décision prise sur le fondement de celui-ci et produisant des effets juridiques à l'égard de l'intéressé ».

Que peut-on faire pour lutter contre les biais ?

« Accroître l'auditabilité des systèmes d'IA » est l'une des voies préconisées par le rapport Villani. À l'heure actuelle, trouver un compromis optimal entre performance et interprétabilité selon le contexte d'application reste du domaine de la recherche.

Cependant, il serait vain de chercher une solution exclusivement technique à un problème qui demeure avant tout politique. Ainsi, l'accès au crédit bancaire est une figure classique des discriminations de genre et d'origine. Aux États-Unis, encore actuellement, l'accès au crédit logement est conditionné par votre lieu de résidence plutôt que par vos capacités de remboursement.

Le rapport Villani propose « la constitution d'un corps d'experts publics assermentés, en mesure de procéder à des audits d'algorithmes, des bases de données et de procéder à des tests par tout moyen requis ». De fait, certaines institutions indépendantes prennent déjà en charge dans leur champ de compétences cet audit public et indépendant, à l'instar de la Cnil et de l'Opesct, même s'il devient opportun de renforcer les capacités d'expertise scientifique et technique sur lesquelles elles pourraient s'appuyer. Par exemple, une meilleure intégration des dispositions législatives à l'arsenal réglementaire constitue l'une des voies de progrès : en France, depuis 2010, pour intégrer les dispositions de la loi du 27 mai 2008 relative à la lutte contre les discriminations, la Cnil a modifié l'autorisation accordée aux banques pour les traitements dits de « credit scoring » attribuant un risque de défaillance et donc un taux différencié d'emprunt aux clients sur la base de leurs profils individuels. En conséquence, les analystes de risque en matière de crédits ne doivent plus prendre en compte depuis cette date le sexe du demandeur de

crédit, pratique jusqu'alors courante mais aboutissant à discriminer les mères de famille monoparentales dans l'accès au crédit.

Il n'empêche : plusieurs fois alertée par certains de ses administrés, la mairie de Villeurbanne, dans le Rhône, lance une opération de « testing » pour objectiver les discriminations qui ont cours dans le secteur bancaire pour l'accès au prêt immobilier et au crédit à la création d'entreprise. Sur l'ensemble de l'agglomération lyonnaise, 90 tests ont été conduits dans 63 agences de 12 banques différentes, d'avril à décembre 2016. Les résultats du testing révèlent des discriminations à chacune des étapes du parcours du demandeur de prêt : ainsi, ils montrent que le genre peut être un critère discriminant dans l'accès au crédit bancaire, mais de façon moins caractérisée que pour l'origine migratoire supposée.

Il devient urgent d'ouvrir les « boîtes noires » de l'IA dans bien d'autres secteurs : une enquête récente menée conjointement par le CNRS et SOS Racisme montre que ces discriminations s'étendent à des domaines qui demeurent insuffisamment explorés en France : la formation professionnelle, l'assurance automobile, l'accès aux complémentaires santé, le crédit à la consommation, l'achat d'une voiture d'occasion, l'hébergement touristique et la reprise d'entreprise.

Signalons à cette occasion, la contribution originale sur l'enjeu sociétal majeur des inégalités dans l'accès à l'emploi réalisée par l'Observatoire des discriminations de l'université Panthéon-Sorbonne qui propose une plateforme internet de différents outils avec un test d'autoévaluation[12].

La formation des professionnels à ces problématiques de discrimination peut également jouer un rôle avec une sensibilisation aux règles éthiques nécessaires pour encadrer les activités basées sur l'IA.

Dans le domaine éthique, l'Ifip a publié un code d'éthique pour les informaticiens professionnels qui codent les programmes et gèrent les systèmes d'information[13]. Même si elles ne possèdent pas la force contraignante des lois, les chartes éthiques servent de

12 https://www.observatoiredesdiscriminations.fr/

13 D. Desbois, « Le code d'éthique et de conduite professionnelle de l'IFIP », *Terminal*, n°128, 2020.

référentiel lorsque le cadre juridique est insuffisant, voire absent pour les firmes transnationales (FTN) gérant des marques et des labels : leur gouvernance étant assez sensible à leur réputation globale du fait de la structure de leur financement et de leurs marchés, les FTN essayent de se conformer aux codes éthiques lorsqu'ils existent, dans une démarche interne le plus souvent préférée à la certification externe.

De fait, le meilleur moyen de se protéger contre ces biais cognitifs est de contraindre les opérateurs de l'IA à ouvrir leurs « boîtes noires » à une expertise publique et indépendante afin de trouver un équilibre durable entre le développement technologique et la protection des citoyens. Ceux-ci doivent être protégés en tant que consommateurs de produits, mais aussi comme producteurs d'interactions sociales dans la sphère du travail autant que dans l'espace public.

L'IA change-t-elle l'emploi ?

En tant qu'administrateur salarié représentant de la CFDT au conseil d'administration d'Inrae, je suis très attentif à cette question dans le contexte de la recherche appliquée et du développement d'innovations pour le secteur agroalimentaire. En tant qu'élu du personnel, l'incidence sur l'emploi est l'une de nos principales préoccupations qui légitime également mes interventions en tant que membre élu du conseil scientifique d'AgroParisTech et observateur CFDT au conseil scientifique d'Inrae.

Il faut bien admettre que, depuis une décennie, l'IA a pris une part de plus en plus importante dans la gestion des processus au sein de l'agroalimentaire pour réduire les coûts, pas seulement en éliminant les gaspillages avec des applications au tri alimentaire et à la gestion des intrants, mais également en contribuant à la réduction de la main-d'œuvre et des coûts qui y sont généralement associés par les gestionnaires, y compris dans la distribution. Cela pose des problèmes de reconversion évidents aux moins qualifiés des salariés que notre syndicalisme doit protéger et accompagner.

Gaia X, le projet pour l'autonomie européenne dans le stockage des données

Des alliances industrielles et des partenariats, en ce cas gouvernemental, Gaia-X est un partenariat franco-allemand visant à combler le retard européen en matière d'informatique de nuage (cloud computing) de façon à assurer une souveraineté minimale sur un marché jusqu'ici dominé par les opérateurs états-uniens comme Amazon, Google et Microsoft ou chinois comme Alibaba. Une plateforme française, Agdatahub, accompagne les opérateurs des filières agricoles pour mettre en œuvre des projets d'aide à la décision

basés sur l'accumulation des données générées par l'Internet des objets de l'agriculture de précision qu'il s'agisse des moissonneuses-batteuses ou des robots de traite.

Les différences d'impact peuvent être sectorielles comme pour les secteurs liés à l'innovation où l'IA est désormais largement utilisée pour le dépôt de brevets[14] ou dans la recherche pharmaceutique.

Pour rassembler les capacités de prévisions et de prospectives, mais aussi faire le lien avec des expérimentations de terrain ciblées sur les catégories d'emploi les plus menacées, le rapport Villani propose la création d'un laboratoire public sur la transformation du travail. Mobilisant un montant global de 32 milliards d'euros, le levier de la formation professionnelle est également évoqué pour développer la médiation numérique et l'innovation sociale afin que l'IA bénéficie à tous. Néanmoins, il convient d'être vigilant afin que cela soit réellement un moyen pour l'individu d'être acteur dans la négociation de sa propre transition professionnelle. De ce point de vue, davantage de recherches socioéconomiques doivent être entreprises sur la réorganisation des chaînes de valeur affectées par l'IA dans le contexte de l'après-Covid-19.

14 V. Kanesarajah, E. White, « Chasing change: Innovation and patent activity during COVID-19. A report on the pandemic's impact on the global R&D community and innovation lifecycle », *Clarivate*, 2020.