Expliquer l'inexplicable ou le droit domptant l'intelligence artificielle

DE STREEL, Alexandre; LOGNOUL, MICHAEL

Published in:
L'Observateur de Bruxelles

Publication date:
2020

Document Version
le PDF de l'éditeur

Dossier spécial

Expliquer l'inexplicable ou le droit domptant l'intelligence artificielle

Alexandre de Streel*
Professeur de droit européen à l'Université de Namur, directeur du Centre de Recherche Information, Droit et Société (CRIDS/NADI) et directeur académique du Centre on Regulation in Europe (CERRE)

Michael Lognoul*
Assistant à l'Université de Namur et doctorant au Centre de Recherche Information, Droit et Société (CRIDS/NADI)
(* Les observations contenues dans cet article appartiennent à leur auteur et n'engagent pas d'autres organismes ou personnes)

I. CONTEXTE ET ENJEUX

L'intelligence artificielle (IA), et plus précisément le *machine learning* (ML), suscite aujourd'hui un engouement sans pareil[1]. Dans la plupart des secteurs de l'économie, des outils fonctionnant grâce à cette technologie sont en cours de développement, ou sont déjà exploités. Qu'il s'agisse de personnaliser des recommandations de produits ou services, de filtrer des contenus illégaux en ligne, de fixer une prime d'assurance, de décider si un crédit peut être alloué, d'établir un diagnostic médical, ou encore de faire rouler une voiture de façon autonome, les systèmes d'intelligence artificielle promettent une efficacité supérieure à celle des êtres humains, pour des coûts moindres.

En bref, le *machine learning* est un algorithme qui fonctionne en apprenant par lui-même quelles règles il va appliquer pour résoudre un problème qui lui est soumis, grâce à des quantités très importantes de données qui lui servent d'exemples. De la sorte, des algorithmes d'apprentissage trouvent des corrélations statistiques au sein de leurs données d'entraînement, et construisent des modèles mathématiques qui en rendent compte. Ces modèles sont ensuite utilisés à des fins de classification ou de prédiction des situations concrètes qui leur sont soumises.

Par exemple, si un outil d'IA est entraîné pour déterminer si un patient souffre d'un cancer, il va déterminer quelles sont les caractéristiques communes aux patients atteints d'une telle maladie, au sein des données d'entraînement qui lui sont soumises, c'est-à-dire de nombreux anciens dossiers médicaux. Par la suite, lorsqu'un nouveau dossier médical d'un patient lui sera soumis à des fins de diagnostic, l'outil ainsi entraîné vérifiera la présence – ou l'absence – des caractéristiques qu'il a associées à la maladie concernée, afin de fournir une réponse à la requête soumise.

Cependant, cette technologie est opaque, à tel point que les outils de *machine learning* sont souvent décrits comme des boîtes noires (*black box*). Cette opacité résulte notamment de la complexité technique des systèmes d'IA. Cela signifie qu'il est difficile, et parfois même impossible, de comprendre comment l'IA parvient à un résultat donné. Même les concepteurs de ces programmes informatiques ne parviennent pas toujours à expliquer comment ceux-ci traitent les informations fournies à l'entrée, pour parvenir aux résultats donnés par ces programmes à la sortie.

L'opacité de ces systèmes de traitement de l'information pose différents problèmes, notamment d'acceptation sociétale et de contestabilité juridique. D'une part, à défaut de pouvoir comprendre comment un résultat déterminé est atteint, il est difficile de faire confiance à la machine qui le produit. D'autre part, les outils d'IA peuvent potentiellement prendre des décisions illégales, par exemple en répliquant des biais présents dans leurs algorithmes ou dans leurs données d'entraînement, comme les discriminations raciales. Ainsi, si une IA est entraînée avec des données relatives à l'octroi de crédits bancaires et si la pratique décisionnelle de la banque dont proviennent les données d'entrainement est entachée de discriminations sur base de la couleur de peau, l'IA répliquera ce biais illégal. Or, l'opacité de ces systèmes rend difficile la contestation en justice et le contrôle juridictionnel des résultats qu'ils fournissent.

1. Sur cette question, voy. l'excellent rapport de la mission de Cédric Villani, *Donner un sens à l'intelligence artificielle : Pour une stratégie nationale et européenne*, mars 2018.

Par conséquent, parvenir à expliquer le fonctionnement et les décisions prises par les systèmes d'IA est un enjeu majeur à l'heure actuelle. Le terme « explication » est ici entendu comme la transmission, aux personnes concernées, d'informations pertinentes et suffisantes sur les motifs, raisons et critères ayant mené à une décision, et sur le fonctionnement du système utilisé pour la prendre. Ainsi, de nombreux chercheurs et industriels actifs dans le domaine de l'IA travaillent déjà sur *l'explainable AI* (*XAI*)[2]. L'explicabilité de l'IA et des décisions prises par le biais de cette technologie font également l'objet d'un nombre croissant de principes éthiques et règles juridiques définis au niveau européen.

II. LE CONSEIL DE L'EUROPE

Le Conseil de l'Europe a modernisé, en 2018, la Convention pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel (Convention 108+)[3]. La Convention reconnaît maintenant à la personne concernée par le traitement de ses données à caractère personnel, le droit d'obtenir connaissance du raisonnement qui sous-tend le traitement de données, lorsque les résultats de ce traitement lui sont appliqués[4].

Le contexte et le contenu de ce droit sont précisés dans le rapport explicatif de la Convention[5]. Celui-ci indique que la nouvelle disposition conventionnelle vise particulièrement l'utilisation d'algorithmes pour adopter des décisions et prend pour exemple l'octroi automatisé d'un crédit bancaire. Dans un tel cas, les emprunteurs ont le droit d'obtenir connaissance de la logique sur laquelle repose le traitement de leurs données et qui aboutit à la décision d'octroi ou de refus du crédit, au lieu d'être simplement informés de la décision elle-même[6]. Le rapport précise également que l'explication du raisonnement qui sous-tend le traitement de données est un prérequis nécessaire pour l'exercice effectif d'un droit de recours contre une décision automatisée.

Par ailleurs, le Conseil de l'Europe a créé en 2019 un comité *ad hoc* sur l'IA (CAHAI). Celui-ci est chargé d'examiner les éléments potentiels d'un cadre juridique international pour l'IA, respectant notamment les droits fondamentaux[7]. A cet égard, la question de l'explicabilité de l'IA fera partie des éléments soumis à la discussion au sein de ce comité.

III. L'UNION EUROPÉENNE

Au niveau de l'Union européenne, la Commission a créé en 2018 un groupe d'experts à haut niveau sur l'IA qui a déjà adopté, d'une part, des lignes directrices proposant des principes éthiques pour une IA digne de confiance et, d'autre part, des recommandations proposant des principes pour développer la régulation et les investissements en IA. Sur cette base, la Commission devrait proposer en février 2020 un livre blanc proposant une approche européenne pour la régulation de l'Intelligence artificielle[8]. Comme l'explicabilité de l'IA est l'un des enjeux principaux liés à la règlementation de cette technologie, il est probable que le livre blanc de la Commission portera notamment sur cette question. Cela étant, le droit de l'Union européenne contient déjà des règles horizontales et sectorielles qui imposent certaines obligations d'explicabilité aux décisions algorithmiques.

A. Groupe d'experts à haut niveau sur l'IA

Le groupe d'expert à haut niveau sur l'IA a adopté en avril 2019 des lignes directrices qui proposent quatre principes éthiques dérivés des droits fondamentaux que devrait respecter tout système d'IA : le respect de l'auto-

2. A cet égard, voy. notamment le projet de la DARPA sur l'*explainable AI* ou les recherches de Google, "Perspectives in Issues in AI Governance", janvier 2019.
3. Convention modernisée 108+ Conseil de l'Europe du 18 mai 2018 pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel.
4. Convention 108+, art. 9(c).
5. Rapport explicatif du 10 octobre 2018 relatif à la Convention modernisée Conseil de l'Europe, pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel.
6. *Ibid.*, p. 15.
7. Décision du Comité des Ministres du Conseil de l'Europe établissant le mandat du Comité ad hoc sur l'intelligence artificielle (CAHAI), CM/Del/Dec(2019)1353/1.5, 11 septembre 2019.
8. Voy. également Ursula von der Leyen, *Une Union plus ambitieuse – Mon programme pour l'Europe*, discours devant le Parlement européen, 16 juillet 2019.

nomie humaine, la prévention des dommages, l'équité et l'explicabilité[9].

Le principe d'explicabilité implique que les processus doivent être transparents, que les capacités et objectifs des systèmes d'IA doivent être communiqués de façon ouverte et que les décisions doivent être, dans la mesure du possible, explicables pour les personnes qui sont affectées directement ou indirectement pour en permettre la contestabilité juridique. Cette explication doit être adaptée au niveau de connaissance des personnes concernées. En outre, le degré d'explication doit être proportionné aux conséquences de la décision algorithmique[10].

Si le modèle algorithmique ne permet pas une explicabilité directe des décisions, ce qui sera souvent le cas des modèles basés sur le *machine learning*, d'autres mesures doivent être prises. Elles peuvent inclure la traçabilité des données qui ont été utilisées pour l'apprentissage de l'algorithme, l'audit de l'algorithme ou la transparence du modèle. Des explications peuvent également être fournies quant à l'intensité du rôle des systèmes automatisés dans les processus de prise de décision, quant aux choix de conception de ces systèmes et quant aux raisons pour lesquelles ils sont déployés. De la sorte, les explications fournies permettront aux parties prenantes de comprendre les décisions prises par IA mais également leur contexte et les modèles d'affaire dans lesquels elles s'intègrent[11].

B. Droit de l'Union européenne

Outre ces principes éthiques, le droit de l'Union européenne contient déjà actuellement des obligations spécifiques en matière d'explicabilité des décisions algorithmiques. Certaines obligations sont prévues par des instruments d'application horizontale, comme le Règlement général sur la protection des données (RGPD)[12], la directive sur les droits des consommateurs[13] ou le règlement sur l'équité et la transparence des services d'intermédiation en ligne[14]. D'autres sont prévues par des instruments d'application sectorielle comme la directive concernant les marchés d'instruments financiers[15].

Le RGPD s'applique en cas de collecte ou de traitement de données personnelles. Lorsque le responsable d'un traitement de pareilles données prend (ou envisage de prendre) une décision fondée exclusivement sur un traitement automatisé, basée notamment sur des données à caractère personnel, et qui produit des effets juridiques ou qui affecte de manière significative la personne concernée, il doit communiquer une série d'informations à la personne concernée par la décision automatisée, avant que celle-ci ne soit prise. Ainsi, il doit informer sur l'existence d'une prise de décision automatisée, y compris un profilage et sur les informations utiles concernant la logique sous-jacente, ainsi que l'importance et les conséquences prévues de ce traitement pour la personne concernée[16]. Dans le même ordre d'idées, le RGPD confère aux personnes concernées par de telles décisions le droit d'obtenir les mêmes informations de la part du responsable du traitement à tout moment, et donc avant et/ou après qu'une décision automatisée ait été prise[17].

De plus, le RGPD prévoit des garanties particulières à la suite de la prise d'une telle décision. Celles-ci portent sur la mise en œuvre, par le responsable du traitement, de mesures appropriées afin de sauvegarder les droits, libertés et intérêts légitimes de la personne visée par la décision automatisée. Ces mesures doivent au moins inclure le droit, pour la personne concernée, d'obtenir une intervention humaine de la part du responsable du

9. High-Level Expert Group on AI, « Ethics Guidelines for Trustworthy AI », 8 avril 2019, p.12.
10. *Ibidem*, p. 13.
11. *Ibidem*, p. 18.
12. Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, J.O., L 119 du 4 mai 2016, p. 1.
13. Directive 2011/83/UE du Parlement européen et du Conseil du 25 octobre 2011 relative aux droits des consommateurs, J.O. L 304 du 22 novembre 2011, p. 64, tel qu'amendée en dernier lieu par la Directive 2019/2161 en ce qui concerne une meilleure application et une modernisation des règles de l'Union en matière de protection des consommateurs.
14. Règlement (UE) 2019/1150 du Parlement européen et du Conseil du 20 juin 2019 promouvant l'équité et la transparence pour les entreprises utilisatrices de services d'intermédiation en ligne, J.O., L 186 du 11 juillet 2019, p. 57.
15. Directive 2014/65/UE du Parlement européen et du Conseil du 15 mai 2014 concernant les marchés d'instruments financiers, *J.O.* L 173 du 12 juin 2014, p. 349 tel qu'amendé, en particulier l'article 17 §2.
16. RGPD, articles 13 §2 (f) et 14 §2 (g).
17. RGPD, article 15 §1 (h).

traitement, ainsi que le droit d'exprimer son point de vue et de contester la décision, ce qui implique le droit d'obtenir une explication de la décision prise [18].

Ces dispositions du RGPD en matière de décisions automatisées ont été interprétées par le Comité européen de la protection des données. Il préconise que le responsable du traitement devrait trouver des moyens d'informer la personne concernée d'une série d'éléments, notamment de la raison d'être de la décision ou des critères sur la base desquels elle a été prise. Le CEPD préconise également que les informations fournies ne doivent pas nécessairement porter sur une explication des algorithmes utilisés mais qu'elles doivent être suffisamment complètes et pertinentes, afin que la personne concernée puisse comprendre les raisons de la décision qui la concerne [19].

Toutefois, le RGPD n'est pas le seul instrument juridique de l'Union à prévoir des obligations en matière d'explicabilité de décisions algorithmiques. La directive relative aux droits des consommateurs a été révisée en 2019 pour imposer des exigences spécifiques supplémentaires en matière d'information applicables aux contrats conclus sur des places de marché en ligne [20]. Elles imposent aux fournisseurs de places de marché en ligne, comme Amazon, Alibaba ou eBay, qui proposent des classement dans les offres présentées aux consommateurs, de fournir des informations sur les principaux paramètres utilisés pour le classement ainsi que l'ordre d'importance de ces paramètres. Ainsi donc, si le classement est effectué par un algorithme, ce qui est presque toujours le cas, les critères sur lesquels se base l'algorithme et leurs pondérations doivent être communiqués aux consommateurs.

Le règlement sur l'équité et la transparence des services d'intermédiation en ligne vient d'imposer une obligation de transparence similaire aux fournisseurs de services d'intermédiation en ligne et de moteurs de recherche en ligne dans leurs relations avec leurs utilisateurs professionnels. Le règlement dispose que ces fournisseurs doivent indiquer les principaux paramètres déterminant le classement et les raisons justifiant l'importance relative de ces principaux paramètres par rapport aux autres paramètres [21]. L'explication à fournir est limitée aux principaux paramètres déterminant le classement et ne s'étend donc pas à tous les paramètres utilisés. Par ailleurs, cette information ne doit pas être spécifique à chaque recherche des utilisateurs, puisque les informations doivent être fournies dans les conditions générales d'utilisation pour les services d'intermédiation, et dans une description accessible au public pour les moteurs de recherche [22]. De plus, le règlement ne requiert pas la divulgation des algorithmes, ni d'aucune autre information susceptible de causer un préjudice aux consommateurs en permettant la manipulation des classements de recherche [23].

IV. CONCLUSION

L'IA est une technologie qui offre de nombreuses opportunités pour la plupart des secteurs de l'économie en améliorant les processus de prise de décision. Toutefois, l'opacité des décisions algorithmiques présente également de nombreux risques, en particulier pour les décisions qui impactent les droits fondamentaux. Il est donc impératif, aux niveaux éthique et juridique, qu'un principe d'explicabilité soit imposé aux décisions algorithmiques, que le droit dompte l'IA et non l'inverse.

Il faut donc se réjouir des initiatives prises récemment au sein du Conseil de l'Europe ou de l'Union européenne pour mieux affirmer ce principe d'explicabilité. Pour le futur, on peut s'attendre à ce que le législateur européen développe dans un instrument juridique plus général sur l'IA une obligation d'explicabilité ou que les juges tentent de construire, à partir des législations existantes et en se basant sur la protection des droits fondamentaux, un principe général d'explicabilité. Cela ne pourra toutefois se faire utilement qu'en créant un dialogue avec les spécialistes de l'IA et en tenant compte des potentialités et des contraintes techniques. Puisse ce dialogue interdisciplinaire être fructueux et mener une explication de ce qui est, pour l'instant, inexplicable.

---

18. RGPD, art. 22, §3 tel que précisé par le considérant 71.
19. Lignes directrices du Groupe de travail Article 29 sur la protection des données du 3 octobre 2017 relatives à la prise de décision individuelle automatisée et au profilage aux fins du règlement 2016/679, révisées le 6 février 2018, WP251rev.01, p. 28.
20. Directive 2011/83/UE tel qu'amendée, article 6 *bis* §1a.
21. Règlement (UE) 2019/1150, art. 5 §§ 1 et 2.
22. *Ibidem.*
23. Règlement 2019/1150, considérant 27.