Explicabilité des algorithmes : à quel niveau faut-il mettre le curseur de la régulation ?

Le néologisme « explicabilité » prend de l'ampleur à l'heure des algorithmes et de l'intelligence artificielle. En Europe, France comprise, la loi exige déjà qu'ils doivent être explicables dans un souci de transparence et de responsabilisation. La question est de savoir s'il faut encore plus réguler.

Par Winston Maxwell* et David Bounie**, Telecom Paris, Institut polytechnique de Paris

L'« explicabilité » est devenue un principe incontournable de la future bonne régulation des algorithmes. L'explicabilité figure dans les recommandations que l'OCDE (1) a adoptées le 22 mai (2) sur l'intelligence artificielle (IA), dans la communication de la Commission européenne du 8 avril (3), dans le rapport du groupe d'experts correspondant (4), et dans le rapport Villani (5). Garante de plus de transparence, l'explicabilité des algorithmes n'est pas un concept nouveau.

Explicabilité : déjà présente dans la loi

Dans les années 1980 et 1990, de nombreux travaux scientifiques en France et aux Etats-Unis ont étudié l'explicabilité de l'intelligence artificielle, explicabilité jugée nécessaire pour promouvoir l'acceptabilité des systèmes « experts » (6). L'année 2018 marque le renouveau du concept. Le rapport Villani évoque la nécessité d'*« ouvrir les boîtes noires »* (7) ; le groupe d'experts européens souligne l'importance de l'explicabilité pour la confiance dans les systèmes IA (8). Que recouvre ce terme qui ne figure nulle part dans les dictionnaires (9) ? Les nouvelles recommandations de l'OCDE font la meilleure synthèse, se focalisant sur les finalités des explications (10). Les législateurs français et européen n'ont pas attendu les conclusions de l'OCDE pour imposer des obligations de transparence.

Le règlement général sur la protection des données (RGPD) impose l'obligation de fournir *« des informations utiles concernant la logique sous-jacente »* de toute décision automatique entraînant un effet important sur la personne (11). La récente révision de la Convention 108 du Conseil de l'Europe (12) confère aux personnes le droit d'obtenir connaissance *« du raisonnement qui sous-tend le traitement »*. La loi « Lemaire » et ses décrets d'application imposent à l'administration l'obligation de communiquer à l'individu les *« règles définissant tout traitement »* algorithmique et les *« principales caractéristiques de sa mise en œuvre »* (14). Ces mêmes textes imposent aux plateformes l'obligation de communiquer les *« critères de classement »*, ainsi que leur *« principaux paramètres »* (15). Au niveau européen, le futur règlement « Plateformes » – surnommé aussi P2B (16) – imposera aux places de marché et aux moteurs de recherche l'obligation d'indiquer les *« principaux paramètres »* qu'ils utilisent pour classer les biens et les services sur leurs sites. Ce règlement, qui devait être adopté le 20 juin 2019 par le Conseil de l'Union européenne, n'imposera pas en revanche la communication de l'algorithme lui-même, protégé par la directive sur le secret des affaires (17). Avec autant de dispositions déjà en place, faut-il encore plus réguler ? L'explicabilité est un problème qui dépasse le débat sur la transparence des plateformes. Elle touche à des secteurs clés de l'industrie, et risque de freiner le déploiement des algorithmes si les contours de l'explicabilité ne sont pas bien définis par le régulateur. La difficulté est à la fois économique et juridique. Sur le plan économique, l'utilisation des algorithmes les plus performants sera freinée si leur fonctionnement n'est pas suffisamment compris par les utilisateurs et les régulateurs. Par exemple, dans la lutte contre le blanchiment d'argent, les algorithmes les plus performants sont également les plus opaques. Or, les banques et leurs régulateurs n'adopteront ces algorithmes que si leur fonctionnement est compris et auditable.

Faire parler ces boîtes noires est donc une condition nécessaire à leur adoption par l'industrie et par l'administration. L'explicabilité est également déterminante pour des questions de responsabilité. En cas d'accident, de discrimination ou toute autre « mauvaise » décision algorithmique, il sera essentiel d'auditer le système afin d'identifier la source de la mauvaise décision, corriger l'erreur pour l'avenir, et déterminer les éventuelles responsabilités. De plus, l'exploitant du système doit être en mesure de démontrer sa conformité avec le RGPD et d'autres textes applicables. Comment démontrer une conformité si le fonctionnement interne de l'algorithme reste énigmatique pour son exploitant ? Une documentation expliquant le fonctionnement de l'algorithme sera nécessaire.

Apprentissage machine et interprétation

Un algorithme d'apprentissage machine *(machine learning)* est l'aboutissement de différentes phases de développement et de tests : on commence par la définition du problème à résoudre, le choix du modèle, le choix des données d'apprentissage, le choix des données de test et de validation,

Notes

(1) - Organisation de coopération et de développement économiques (OCDE).

(2) - https://lc.cx/IA-OCDE22-05-19

(3) - https://lc.cx/IA-CE08-04-19

(4) - https://lc.cx/AI-Guidelines08-04-19

(5) - Rapport Villani 28-03-18 : lc.cx/Reso

(6) - On utilisait également le terme de l'explicabilité des systèmes d'expert.

(7) - Rapport Villani, p. 21.

(8) - Rapport du groupe d'experts, p. 13.

(9) - Le mot explicabilité *(explainability)* ne figure ni dans le dictionnaire de l'Académie française, ni dans les dictionnaires américains.

les ajustements des paramètres *(tuning)*, et le choix des données d'exploitation. Le modèle qui émerge après son apprentissage sera très performant mais presqu'aussi inscrutable que le cerveau humain. Si les théories mathématiques sous-jacentes aux modèles utilisés sont bien comprises, il est délicat – pour ne pas dire souvent impossible – de comprendre le fonctionnement interne de certains modèles. C'est le cas bien souvent de certains modèles tels que les machines à vecteurs de support, les forêts aléatoires, les arbres améliorés par gradient, et les algorithmes d'apprentissage profonds tels que les réseaux de neurones artificiels, les réseaux de neurones convolutifs et les réseaux de neurones récurrents difficiles à interpréter.

Transparence, auditabilité et explicabilité

Le concept d'explicabilité – ou « explainable AI » (XAI) en anglais) –, parfois désigné par intelligibilité, est un thème de recherche en informatique en plein essor. Il est en particulier soutenu par un programme ambitieux de l'agence du département de la Défense des Etats-Unis, Darpa (18). Les recherches s'orientent sur le développement de méthodes qui aident à mieux comprendre ce que le modèle a appris, ainsi que des techniques pour expliquer les prédictions individuelles. Les solutions techniques se regroupent autour de deux familles : l'identification des facteurs les plus importants utilisés par le modèle (« *saliency approach* »), et la perturbation des données d'entrée afin de comprendre l'impact sur les décisions de sortie (« *perturbation approach* »). Ces approches permettront la visualisation des facteurs déterminants dans la prise de décision algorithmique.

Aux concepts d'interprétabilité et d'explicabilité sont associés ceux de transparence et d'« auditabilité » des algorithmes. L'idée est de rendre publics, ou bien de mettre sous séquestre, des algorithmes en vue les auditer pour étudier des difficultés potentielles. Comme illustré dans le débat sur le règlement « Plateformes » (P2B), imposer la communication des algorithmes se heurterait à la protection des secrets d'affaires, protégée par la directive européenne du 8 juin 2016 (19). Cependant, il existe des solutions intermédiaires, qui permettraient à la fois un audit approfondi du système tout en protégeant le secret des affaires. La Cnil recommande de mettre en place une plateforme nationale d'audit des algorithmes (20). L'une des difficultés pour le régulateur est l'impossibilité de répliquer le fonctionnement d'un algorithme à un moment T. Un algorithme d'apprentissage machine peut se mettre à jour régulièrement, apprenant à améliorer sa performance au fur et à mesure, en fonction des nouvelles données analysées. Ainsi, l'algorithme pourra donner un score de 93,87 % le lundi, et un score de 94,28 % le jeudi, s'appuyant sur exactement les mêmes données d'entrée. Cette absence de reproductibilité peut poser problème pour un régulateur en charge de valider un système avant sa mise en service, car le modèle changera lors de chaque nouvelle phase d'apprentissage. Pour un algorithme de moteur de recherche, ce n'est pas grave. Mais pour un algorithme pouvant provoquer des dommages corporels en cas d'erreur, cette évolutivité posera problème.

Exiger une transparence totale sur le fonctionnement de l'algorithme ne servirait à rien. Une telle exigence se heurterait à la protection du secret des affaires et, dans la plupart des cas, ne rendrait pas l'algorithme plus intelligible, même pour les initiés. La réglementation doit plutôt se concentrer sur les différentes finalités de l'explicabilité et les différents publics visés – grand public, régulateur, expert judiciaire. Une explication envers un demandeur de crédit pourrait privilégier une approche dite « contrefactuelle », par laquelle l'individu pourrait tester d'autres hypothèses d'entrée pour voir l'effet sur son score algorithmique.

Les outils d'explicabilité envers un régulateur seraient plus élaborés, permettant une explication du fonctionnement global de l'algorithme (dit « *global explainability* »), ainsi qu'une explication concernant chaque décision particulière (« *local explainability* »). Les outils mis à la disposition du régulateur pourraient inclure des cartographies de pertinence *(saliency maps)*, et autres outils de visualisation statistique. Enfin, en cas d'audit, l'exploitant de l'outil devra être en mesure de donner accès à l'ensemble de la documentation concernant l'algorithme et les données d'apprentissage, et permettre au régulateur de conduire des tests, notamment des tests de perturbation de données d'entrée.

L'explicabilité, comme la régulation, est source de coûts. Un fort niveau d'explicabilité peut limiter le type de modèle utilisé, conduisant à des pertes de performance. Cependant, certains estiment que le compromis entre performance et explicabilité n'est que temporaire, et disparaîtra avec le progrès technique. La régulation est également source de coûts, surtout lorsque la régulation ne suit pas l'évolution technologique.

Privilégier une régulation « *bacs à sable* »

Ainsi, la Commission européenne et le Conseil d'Etat recommandent la régulation expérimentale – des « *bacs à sable* » de régulation – permettant de tester différentes approches de régulation pour déterminer celle qui sera la plus efficace. Le principe européen de la proportionnalité guidera le niveau de régulation. Le niveau d'exigences en matière d'explicabilité augmentera avec le niveau de risques. Pour certains usages, aucune explicabilité ne sera nécessaire. Pour un algorithme qui apprend à jouer aux échecs, le fonctionnement interne ne soulève pas d'enjeu réglementaire. Par conséquent, la boîte noire pourra rester fermée. @

** Winston Maxwell, ancien avocat associé du cabinet Hogan Lovells, est depuis juin 2019 directeur d'étude, droit et numérique à Telecom Paris, Institut polytechnique de Paris.*

*** David Bounie est professeur d'économie à Telecom Paris, Institut polytechnique de Paris.*

Notes

(10) - Point 1.3 « Transparence et explicabilité » des recommandations de l'OCDE.

(11) - Article 13, RGPD.

(12) - https://lc.cx/Convention108

(13) - Article 9(1)(c) de la Convention 108.

(14) - Article 4, loi du 7 octobre 2016 pour une République numérique.

(15) - Décret n° 2017-1434 Du 29 septembre 2017.

(16) - https://lc.cx/EquitéTransparence

(17) - Directive 2016/943.

(18) - Un programme similaire a été lancé par le département de la défense américaine dans les années 1980 sous l'acronyme EES (Explainable Expert Systems).

(19) - Directive « Secret des affaires » transposée en France par la loi du 30 juillet 2018.

(20) - https://lc.cx/Cnil-Ethique