RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

IGPDE

ACTION PUBLIQUE

Recherche et pratiques

Intelligence artificielle : le secteur public à l'aube d'une révolution ?

Décembre 2024
2024/4
N° 23

ACTION PUBLIQUE

N° 23

Recherche
et pratiques

Directrice de la publication
Marie Niedergang

Rédactrice en chef
Marie Ruault

Co-rédacteur en chef
Edoardo Ferlazzo

Secrétaire de rédaction
Delphine Mantienne

Assistante de rédaction
Louise Langlois

Design graphique
Studio graphique des ministères économiques et financiers (SG-Sircom)

Mise en page et maquettage
Desk

Publication trimestrielle en accès libre – ISSN 2647-3135 (en ligne)

Si vous souhaitez consulter les numéros de la revue *Action publique. Recherche et pratiques*, rendez-vous sur le site https://www.cairn.info/revue-action-publique-recherche-et-pratiques.htm.

Pour vous inscrire à notre liste de diffusion : recherche.igpde@finances.gouv.fr

Réseaux sociaux :
Twitter : @Igpde_recherche
Youtube.com/igpde

S'abonner à la revue :

CAIRN.INFO
MATIÈRES À RÉFLEXION

Sommaire

Intelligence artificielle : le secteur public à l'aube d'une révolution ?

Éditorial

MARIE NIEDERGANG

Marie Niedergang
Directrice générale
de l'Institut
de la gestion publique
et du développement
économique

Les 23es Rencontres internationales de la gestion publique (RIGP) ont eu lieu à Bercy le 14 novembre 2024 sur le thème « Gouverner (par) l'IA : l'action publique à la croisée des chemins ». Cette édition a montré dans quelle mesure l'intégration dans le quotidien administratif de solutions technologiques issues de l'IA était susceptible de transformer, parfois de manière assez profonde, le pilotage, la conduite et l'évaluation de l'action publique (*gouverner par l'IA*) et nombre de ses dimensions : la complexité administrative, le contrôle et la lutte contre la fraude, la distribution des prestations sociales et des services publics.... Au-delà, elle a souligné que l'IA, comme tout outil, présente des risques qu'il s'agit de réguler (*gouverner l'IA*), à des niveaux techniques (sécurité des données, maintenance...), politiques (souveraineté, protection des données individuelles...), économiques et sociaux (création d'inégalités « algorithmiques », suppressions d'emplois...) mais aussi éthiques et culturels (socialisation et formation des agents à une nouvelle culture numérique, adaptation des cultures métier...). Ce numéro propose d'approfondir et d'enrichir ces RIGP.

Le *Regards croisés* fait dialoguer **Thomas Cottinet**, directeur de l'Ecolab aux ministères Aménagement du territoire Transition écologique, et **Thomas Le Goff**, maître de conférences en droit et régulation du numérique à Télécom Paris/Institut polytechnique de Paris, à propos des dimensions environnementales de l'IA. Un premier article de **Clément Henin**, directeur du service de science des données à l'Assistance publique – Hôpitaux de Paris et auteur d'une thèse en informatique, aborde les enjeux d'explicabilité du fonctionnement de l'IA au sein du travail administratif. Un second article de **Winston Maxwell**, avocat aux barreaux de New York et Paris et directeur d'études Droit et Numérique à Télécom Paris/Institut polytechnique de Paris, investit la régulation de l'IA aux États-Unis et ses conséquences en matière de décision administrative. Enfin, la *Note réactive* d'**Amicie de Tanoüarn** présente le programme de formation à l'IA avant-gardiste des fonctionnaires irlandais.

Excellente lecture !

Introduction

Le comité de rédaction

L'histoire de l'IA a pris racine dans les années quarante et le développement de la cybernétique, alors même que le terme d'intelligence artificielle n'existait pas encore. Elle a vu se succéder des controverses entre deux courants scientifiques, l'IA connexionniste et l'IA symbolique, dont les promesses, les progrès mais aussi les espoirs déçus ont alimenté plus généralement les débats sur la place de l'IA dans la société au cours des dernières décennies (Cardon *et al.*, 2018). Les développements que l'IA – notamment connexionniste et liée aux réseaux de neurones – a connus depuis une dizaine d'années, rendus possibles par la montée en puissance de microprocesseurs plus performants et par la quantité exponentielle de données à disposition des entreprises informatiques (*big data*), ont néanmoins fait émerger des résultats technologiques spectaculaires, sans commune mesure avec les expériences précédentes. L'essor est tel que certains observateurs voient dans la période actuelle l'aube de la quatrième révolution industrielle.

Pourtant, il n'existe aucune définition de l'IA qui fasse consensus. Elle peut par exemple être assimilée à un ensemble de produits ou de services utilisant un ou des programmes informatiques issus de l'IA. Ces produits peuvent être caractérisés en fonction de leur finalité, de leurs fonctionnalités (prévision, traduction, reconnaissance…) ou du type de modèle qu'ils utilisent (apprentissage automatique, système expert…). Loin des fantasmes qui imaginent déjà une « IA forte », c'est-à-dire capable de répliquer de façon autonome le comportement humain, ils concernent essentiellement une « IA faible », c'est-à-dire reproduisant certaines tâches spécifiques, plus ou moins complexes, associées aux fonctions cognitives de l'humain (parler, écouter, percevoir…), grâce à une assistance humaine pour les concevoir ou les faire fonctionner. Il est néanmoins difficile de catégoriser de manière claire l'ensemble des objets concernés par l'IA, tant les entités techniques qu'elle désigne tendent à évoluer rapidement dans le temps (Benbouzid et Cardon, 2022).

D'autre part, et sans doute en premier lieu, l'IA est, un champ de recherche qui mêle des sciences dures, en mathématiques et en informatique, et des sciences sociales, en linguistique, philosophie, sociologie, économie, droit et éthique (Vayre et Gaglio, 2020). Ces travaux académiques alimentent la recherche fondamentale, permettant à la fois d'atteindre certains progrès scientifiques et de produire des innovations technologiques mises sur le marché.

Les redéfinitions permanentes de ce que constitue l'IA ont des conséquences importantes sur la manière dont le débat public s'en empare et surtout sur la manière dont chaque catégorie d'acteurs concernée souhaite la réguler (Benbouzid et Cardon, 2022). Comme le suggèrent Benbouzid *et al.* (2022) à propos des débats autour des enjeux de régulation de l'IA, « les problèmes définitionnels [sont] au cœur de conflits normatifs sur les moyens d'assujettir l'IA à un "contrôle social", qu'il soit technique, éthique, juridique ou politique ». Deux conséquences majeures qui influent sur la régulation de l'IA surgissent de ces enjeux de définition.

D'une part, compte tenu du nombre d'acteurs impliqués (entreprises du numérique, chercheurs en IA, associations, États…), chacun tend actuellement à proposer des réflexions et des mesures concrètes de régulation (charte, lois…) qui interrogent la capacité des décideurs publics à la fois à intégrer des solutions IA pour élaborer, déployer et évaluer l'action publique, mais aussi, et plus généralement, à en réguler la diffusion au sein des sociétés.

D'autre part, face aux mutations rapides et difficiles à anticiper que connaît le champ de l'IA, sa régulation se heurte à ce que Collinridge (1980) appelle le dilemme du « contrôle social des technologies », caractérisé par un manque de connaissances pour prédire les conséquences de ces technologies dans leur phase d'essor. Autrement dit, comme le précisent Benbouzid et Cardon (2022), « nous sommes contraints de les laisser se déployer pour en mesurer les conséquences, mais il est alors trop tard car elles sont enracinées dans la société et par un effet de dépendance leur contrôle devient difficile ». Confrontés à cet essor spectaculaire de l'IA lors des dernières années, les pouvoirs publics n'ont pas manqué de prendre le sujet à bras le corps en tentant de réguler les usages de l'IA, comme en témoigne l'AI Act européen.

Ce numéro de la revue étudie certains enjeux qui jalonnent le déploiement de l'IA et sa difficile régulation par et pour le secteur public, en montrant notamment que tout cela se joue entre plusieurs échelons d'action.

Tout d'abord, la régulation de l'IA questionne son contrôle social et, par là même, ses effets sociétaux et environnementaux. L'impact écologique de l'IA apparaît à ce titre comme un enjeu des plus controversés. Là où l'IA est fréquemment présentée comme un moyen de répondre à la crise écologique, elle génère dans le même temps une empreinte environnementale considérable, notamment en matière d'émissions de CO_2 et de consommation d'eau (Le Goff, 2023). S'il existe des solutions techniques pour limiter cette empreinte (processeurs plus performants, optimisation énergétique des centres de données…), leur mise en œuvre est tributaire de choix politiques, économiques et juridiques, ce qui plaide pour une régulation transnationale.

Ensuite, si les réglementations nationales ou supranationales ont fleuri ces derniers mois, la question demeure de l'application de leurs principes généraux dans des contextes plus locaux. En radiologie, le champ pionnier où la question de la régulation des outils IA s'est posée, Mignot et Schultz (2022) ont montré qu'en dépit de principes abstraits présents dans des chartes et des rapports, la régulation s'est opérée au niveau des acteurs du domaine, à savoir les radiologues et les industriels du secteur. En conséquence, la régulation qui a émergé s'est d'abord construite « autour des délimitations du groupe professionnel des radiologues et de la compétition entre les constructeurs historiques de dispositifs d'imagerie et les nouveaux entrants de l'innovation numérique » (Mignot et Schultz, 2022). En s'alignant sur les principes généraux, les utilisateurs de l'IA les interprètent localement, à l'échelle d'un ministère, d'une branche, d'un métier, d'une direction ou d'un service par des normes de droit plus « molles » (décrets, circulaires…). C'est aussi dans son adaptation à l'échelle du collectif de travail que la régulation de l'IA trouvera son expression.

Enfin, compte tenu de la démultiplication des cas d'usage au sein de l'administration, la régulation prend parfois la forme d'un contrôle au plus près de la machine. Une autre échelle de régulation ne se situerait-elle pas dans le périmètre des utilisateurs, c'est-à-dire à l'échelle de celles et ceux ayant à voir avec les décisions prises par l'IA, soit qu'ils conçoivent les machines et en contrôlent les performances techniques, soit qu'ils les utilisent comme aide à la décision ou à l'information dans leur travail administratif, soit qu'ils en subissent les décisions, en tant qu'usagers du service public ? Pour reprendre les mots de Callon (1986), au sein de ce réseau d'innovation, tous les acteurs – bien qu'ayant des intérêts divergents – doivent être enrôlés autour de l'IA pour que cette dernière soit légitimée. Les enjeux de transparence et d'explicabilité, c'est-à-dire la manière dont les machines prennent leurs décisions, deviennent alors décisifs pour procéder à des opérations de traduction qui explicitent et, potentiellement, alignent les positions de chacun. Il s'agit en ce sens de ménager des possibilités, pour les autorités de contrôle, pour les agents publics, mais aussi pour les usagers du service public augmenté par l'IA, de justifier ou contester une décision prise par l'IA. Cette explicabilité est garante de la crédibilité et de l'impartialité de la machine pour ses usagers-humains. Elle permet à la fois une délimitation claire des responsabilités entre humain et machine et la légitimation de cette dernière pour améliorer le travail administratif. À ce titre, le gouvernement britannique a lancé à l'automne 2021 un des premiers standards nationaux de transparence algorithmique pour les organisations publiques. Construit à partir de délibérations publiques et à l'issue de réunions de fonctionnaires de plusieurs pays, de chercheurs et de spécialistes du secteur, ce standard vise à apporter des solutions concrètes pour expliquer au public les processus décisionnels et pour comprendre les modalités de l'utilisation des algorithmes dans l'action publique.

En somme, ce numéro révèle que le déploiement et la régulation de l'IA se jouent au creuset d'inter-dépendances entre niveaux local et global. À l'heure où de nombreuses expérimentations d'IA sont déployées au cœur des administrations françaises et étrangères, nul doute que les retours du terrain, issus des agents publics eux-mêmes ou des analyses des chercheurs, ne manqueront pas de faire évoluer cette régulation au rythme des usages.

Références

Benbouzid B., Meneceur Y. et Smuha N. (2022),
« Quatre nuances de régulation de l'intelligence artificielle. Une cartographie des conflits de définition », *Réseaux*, 2022/2, nº 232-233 « Contrôler l'intelligence artificielle ? », pp. 29-64. https://doi.org/10.3917/res.232.0029.

Benbouzid B. et Cardon D. (2022),
« Contrôler les IA », *Réseaux*, 2022/2, nº 232-233 « Contrôler l'intelligence artificielle ? », pp. 9-26. https://doi.org/10.3917/res.232.0009

Callon M. (1986),
« Éléments pour une sociologie de la traduction. La domestication des coquilles Saint-Jacques et des marins pêcheurs dans la Baie de Saint-Brieuc », *L'Année sociologique*, nº 36, pp. 169-207.

Cardon D., Cointet J. et Mazières A. (2018),
« La revanche des neurones. L'invention des machines inductives et la controverse de l'intelligence artificielle », *Réseaux*, 2018/5, nº 211 « Machines prédictives », pp. 173-220. https://doi.org/10.3917/res.211.0173

Collinridge D. (1980),
The Social Control of Technology, New York, St. Martin's Press.

Le Goff T. (2023),
« Recommandations pour une action publique en faveur d'une IA générative respectueuse de l'environnement », document de travail. https://hal.science/hal-04371031

Mignot L. et Schultz É. (2022),
« Les innovations d'intelligence artificielle en radiologie à l'épreuve des régulations du système de santé », *Réseaux*, 2022/2, nº 232-233 « Contrôler l'intelligence artificielle ? », pp. 65-97. https://doi.org/10.3917/res.232.0065

Vayre J. et Gaglio G. (2020),
« L'intelligence artificielle n'existe-t-elle vraiment pas ? Quelques éléments de clarification autour d'une science controversée », *Diogène*, 2020/1, nº 269-270, pp. 107-120. https://doi.org/10.3917/dio.269.0107

Regards croisés

Entre recherche et pratiques

Les *Regards croisés* reposent sur un dialogue organisé entre une personne issue du monde académique et universitaire et une personne issue de l'administration publique sur un sujet d'intérêt commun.

Ce dialogue est animé dans le cadre d'une interview vidéo publiée sur la chaîne YouTube de l'IGPDE. Cette interview est également retranscrite et remaniée sous la forme d'un article publié dans cette revue.

La régulation au cœur des enjeux d'une IA frugale ?

Entretien entre Thomas Cottinet et Thomas Le Goff[1]

Thomas Cottinet est directeur de l'Ecolab, laboratoire d'innovation au service de la transition écologique dépendant du Commissariat général du développement durable (CGDD), acteur interministériel et direction transversale des ministères Aménagement du territoire Transition écologique.

Thomas Le Goff est maître de conférences en droit et régulation du numérique à Télécom Paris – Institut polytechnique de Paris (laboratoire i3, CNRS, UMR 9217).

Retrouvez cet entretien en vidéo sur le site de la revue

www.economie.gouv.fr/igpde-editions-publications/action-publique-recherche-pratiques

1 Cet entretien, animé par Marie Ruault, directrice de la Recherche à l'IGPDE, a été enregistré le 29 novembre 2024.

L'usage de l'intelligence artificielle peut-il être justifié dans le cadre de la lutte contre le réchauffement climatique ?

Thomas Le Goff : On peut tout à fait le dire. C'est un phénomène qui est plutôt bien documenté dans la doctrine. Depuis plusieurs années, des études essaient d'évaluer le potentiel de l'intelligence artificielle au service des objectifs de développement durable. Par exemple, une grande étude a été publiée en 2019 dans la revue *Nature*[2], examinant comment l'IA pouvait être mise au service des 15 objectifs de développement durable de l'ONU. D'autres études se concentrent davantage sur le réchauffement climatique, comme celle d'un groupe de chercheurs, en 2019[3], qui a essayé de recenser tous les usages de l'IA dans les grands secteurs d'activité comme l'énergie ou l'agriculture. Peut-être aurons-nous par la suite des exemples plus concrets de la façon dont l'IA peut contribuer à décarboner ces secteurs-là. Avant d'être chercheur, j'ai travaillé dans le secteur de l'énergie. J'ai donc quelques exemples en tête. En effet, toutes les activités du secteur de l'énergie peuvent être concernées par ce potentiel de l'IA en faveur de la transition énergétique, à la fois dans la production, avec des usages de l'IA pour faire de la maintenance prédictive, pour mieux prédire la production des énergies renouvelables ; dans la gestion des réseaux pour l'optimiser et faciliter l'intégration des énergies renouvelables ; ou même dans les services énergétiques aux personnes avec des solutions pour analyser les consommations d'énergie et faire des recommandations d'économie d'énergie. C'est un sujet qui commence à être bien documenté dans le secteur de la recherche et on peut également observer un certain nombre d'initiatives industrielles, c'est donc un potentiel non négligeable. Je mettrais néanmoins un bémol : aujourd'hui, dans l'état de l'art de la recherche, nous ne sommes pas en capacité de mesurer le potentiel global de l'IA dans ses usages « positifs ». Aucune étude transsectorielle ne mesure ce potentiel, parce qu'il y a trop de paramètres à prendre en compte. Il y a des questions de passage à l'échelle de la technologie. On a beaucoup de cas d'usage très prometteurs qui sont très bien documentés sur leur potentiel individuel, mais l'impact global dépend de trop de facteurs différents pour savoir si le potentiel positif va réellement compenser les effets négatifs. Ce manque de recherches sur le potentiel global est un verrou scientifique assez important, puisqu'il conditionne ensuite toutes les réflexions sur la régulation et les autres mesures pour limiter les impacts environnementaux. Je donne un exemple très concret. DeepMind, il y a un certain temps, avait publié un article disant que ses chercheurs avaient développé un algorithme pour augmenter de 40 à 50 % l'efficacité énergétique des *data centers*[4]. Néanmoins, ils reconnaissaient eux-mêmes qu'ils ne pouvaient l'appliquer que dans un des *data centers* de Google, parce que cela nécessitait un *design* très spécifique. Aujourd'hui, on a avancé par rapport à cela, mais c'est un exemple. On entend de plus en plus de discours du type : « Très bien, on a un cas d'usage qui est vraiment très prometteur et qui va nous permettre de réaliser beaucoup d'économies », mais on a encore peu de données sur le passage à l'échelle et les impacts que peut avoir l'adoption de la technologie sur toute une filière. Certains secteurs d'activité se sont malgré tout structurés avec des feuilles de route, etc., pour utiliser l'IA afin de décarboner leur activité. Mais aujourd'hui, les cas d'usage arrivent plutôt par le bas.

L'Ecolab du ministère de la Transition écologique accompagne des projets numériques, notamment en matière d'IA. Pouvez-vous nous présenter quelques usages emblématiques qui permettent de lutter contre le réchauffement climatique ?

Thomas Cottinet : Nous devons éviter l'émission de 138 millions de tonnes de CO_2 en seulement 6 ans. C'est plus que le total des émissions évitées

2 Vinuesa R. *et al.* (2020), "The role of artificial intelligence in achieving the Sustainable Development Goals", *Nature Communications*, 11, 233, https://www.nature.com/articles/s41467-019-14108-y.
3 Rolnick D. *et al.* (2019), "Tackling Climate Change with Machine Learning", https://arxiv.org/pdf/1906.05433.
4 DeepMind (2016), "DeepMind AI Reduces Google Data Centre Cooling Bill by 40%", https://deepmind.google/discover/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-by-40/.

lors des 33 dernières années cumulées. Le trait de côte français recule sur 55 % du littoral français. Il y a deux ans, 90 départements sur 100 ont été concernés par de grands incendies de forêt, ce qui n'était jamais arrivé. On a connu une sécheresse hivernale très longue il y a un an et demi ; 70 % de la faune sauvage a disparu. Qu'il s'agisse d'atténuation face aux émissions ou d'adaptation au changement climatique, le défi est vertigineux. Nous ne sommes pas dans le solutionnisme technologique. L'intelligence artificielle n'est pas la baguette magique, mais elle est utile, en particulier pour *accélérer*, parce qu'il faut pouvoir faire plus vite et plus fort. Elle est utile, mais pas n'importe comment. Quelques exemples de cas d'usage. En premier lieu, ceux-ci viennent de « quelque part » : il faut pouvoir les faire émerger. La France est dotée d'une feuille de route « Intelligence artificielle pour la transition écologique ». Une communauté d'acteurs publics et privés s'est créée, qui mêle territoires, entreprises et chercheurs pour stimuler et accroître le recours effectif à l'intelligence artificielle pour la transformation écologique des territoires. On peut citer plusieurs exemples. Tout d'abord, l'amélioration du pilotage énergétique d'un très important parc de 200 bâtiments publics à Noisy-le-Grand grâce à l'intelligence artificielle. Ensuite, de manière plus générale, une meilleure organisation des rénovations des constructions. Où faut-il rénover ? Comment ? Où faut-il construire ? En métropole de Bordeaux, l'intelligence artificielle sert à planifier cela. Comment élaborer des plans locaux d'urbanisme qui soient plus favorables à la transition écologique ? Comment tester des scénarios d'urbanisation ? Il existe sur ces questions un projet porté à Saclay qui lui aussi utilise l'intelligence artificielle. Comment accélérer et être plus efficace sur la végétalisation des territoires ? La métropole de Lyon tente de répondre à ces problématiques en jouant à la fois sur la plantabilité, sur le rafraîchissement et sur la désimperméabilisation des sols. Là aussi, l'intelligence artificielle est extrêmement utile dans son potentiel d'exploitation des données. Ou encore pour améliorer le nettoyage, le ramassage des déchets. En métropole de Metz, on utilise toutes les données captées par les personnes qui sont en charge de la propreté du territoire. Les cas d'usage sont donc multiples. Il y en a également un grand nombre sur des sujets où est en jeu une forme de pluridisciplinarité. Par exemple, associer l'enjeu de la mobilité durable avec celui de la pollution de l'air ou du bruit. Ainsi, en Île-de-France, un consortium associe l'étude des flux de mobilité avec celle de la qualité de l'air sur l'ensemble du territoire. Dans un territoire plus particulier d'Île-de-France, le Val-de-Marne, sont menés des tests qui mixent les problématiques relatives au bruit, à la qualité de l'air et aux mobilités, pour faire de meilleurs choix au niveau territorial en prenant en compte ces trois enjeux. L'intelligence artificielle associée à la mobilisation de la donnée permet une nouvelle façon de faire de l'aide à la décision, de la prédiction, de la classification, de l'optimisation. Elle rend aussi possible une meilleure catégorisation des enjeux.

Comment l'action publique peut-elle orienter l'innovation vers une IA plus durable ?

Thomas Le Goff : Il est vraiment essentiel que l'action publique promeuve ces usages vertueux de l'intelligence artificielle. Et au-delà de ce que fait l'Ecolab sur l'accompagnement des jeunes pousses, sur la mise en visibilité, etc., si je parle de nos établissements publics d'enseignement et de recherche, je pense qu'il est très important d'intégrer dans la formation de nos ingénieurs des bases sur la transition sociale écologique, mais aussi des bases théoriques pour comprendre comment l'ingénieur est acteur de cette transition, et comment il peut développer des solutions d'intérêt public. C'est quelque chose que nous avons commencé à faire à Télécom Paris. L'école a créé l'an dernier, avec beaucoup de collègues de différentes spécialités, un cours de transition sociale et écologique, qui est un gros bloc dispensé sur l'année entière. Je pense que le levier de la formation est très important, non seulement dans l'enseignement supérieur et le milieu de l'ingénierie, mais pour toutes les disciplines et pour tous les niveaux de formation. La sensibilisation doit commencer le plus tôt possible, parce que si l'on est plutôt aligné sur les causes du réchauffement climatique et les solutions à mettre en place, la réalité, c'est qu'il faut diffuser cette conscience écologique au plus grand nombre, et notamment aux acteurs du monde technologique de demain. Au-delà du levier pédagogique, je dirais qu'il y a aussi le levier du financement, évidemment. La recherche a besoin d'argent, de fonds publics pour mener des projets à visée d'intérêt général. Car évidemment, quand on court après de l'argent privé, on est parfois conditionnés par des intérêts ou par des objectifs de recherche qui ne sont pas forcément les intérêts publics. Même si on commence à observer un alignement, la réalité, c'est que le financement public de la recherche a encore

un rôle très important aujourd'hui, qui va se développer à l'avenir. Or, on sait qu'en France, ce financement n'est pas très élevé comparativement à certains autres grands pays comme les États-Unis. Ce levier du financement passe par des appels à projets fléchés ou par la prise en compte de critères environnementaux dans les appels à projets d'IA, par exemple. C'est d'ailleurs une pratique qui commence à être développée dans les grands congrès internationaux de recherche sur l'intelligence artificielle, où il y a maintenant presque systématiquement une sous-question sur l'empreinte de la solution proposée. Ceci est vraiment essentiel, car aujourd'hui les recherches en IA qui sont les plus primées impliquent dans la plupart des cas la publication de gros modèles dont l'impact environnemental n'est pas négligeable. Flécher l'investissement dans cette direction, mais aussi construire et promouvoir des partenariats publics et privés me semble essentiel. Enfin, je voudrais évoquer la question du levier de la régulation, dans lequel l'action publique a un rôle à jouer. La régulation, c'est un sujet important pour alléger dès que possible les contraintes réglementaires pesant sur des cas d'usage qui ont une forte valeur ajoutée en termes d'intérêt social pour lutter contre le réchauffement climatique. Nous devons mener la réflexion de façon collaborative, nous, membre des communautés de recherche, mais c'est aussi le rôle des pouvoirs publics d'identifier tous les verrous réglementaires – par exemple, liés à l'accès aux données – disproportionnés, injustifiés ou qui pourraient être allégés pour permettre un certain nombre de cas d'usage. On peut également imaginer la mise en place de bacs à sable réglementaires, un dispositif qu'on commence à connaître plutôt bien. Il y en a dans le secteur de l'énergie, qui sont pilotés par la commission de régulation de l'énergie, mais aussi en application du règlement général à la protection des données, qui sont pilotés par la Cnil. C'est à mon avis un dispositif très prometteur, parce qu'il permet à des *startups* de se développer dans un environnement moins contraignant et avec un accompagnement renforcé. Et si on publie les résultats en étant transparent sur les leçons qu'on a apprises, on en fait bénéficier tout l'écosystème. Cet instrument commence aujourd'hui à être utilisé un peu partout et a devant lui un avenir très important. Outre l'allègement des contraintes, la régulation a aussi pour rôle de flécher l'innovation en essayant de détourner la recherche et le développement de systèmes d'IA qui ne seraient pas durables. Il nous faut avoir une régulation graduée en fonction de l'intérêt social, car il y a du potentiel positif pour la lutte contre le réchauffement climatique.

Quel rôle l'Ecolab joue-t-il en matière de réglementation et de régulation de l'IA pour la faire tendre vers le respect des objectifs environnementaux ?

Thomas Cottinet : Nous travaillons sous trois angles. Premièrement, et nous venons de l'évoquer : comment l'intelligence artificielle peut-elle aider et accélérer la transition écologique ? Deuxièmement : comment peut-elle contribuer à moderniser les administrations qui portent ces politiques publiques ? Et troisièmement : comment peut-on faire tout cela dans un cadre frugal ? Le sujet transverse, quand on essaie d'avancer sur ces thèmes-là, c'est la donnée. C'est-à-dire que rien n'est possible sans un accès effectif à la donnée et à la connaissance. Une première brique très importante est le travail à mener sur la « découvrabilité », la « trouvabilité » et la qualité de cette donnée qui va ensuite être exploitée par l'intelligence artificielle. La semaine dernière, nous avons rendu publique la plateforme ecologie.data.gouv.fr. Ce sont 30 000 bases de données avec des métadonnées qualifiées, et c'est un véritable terreau à l'innovation, qui est tout aussi important que le reste. Il me semble que l'on a vraiment besoin de spécifier des cas d'usage : que l'on soit sur des enjeux de biodiversité, d'énergie, d'économie circulaire, de santé-environnement, de production d'énergie renouvelable, de mobilité durable, à chaque fois, il y a un écosystème complètement différent d'acteurs publics, privés, territoriaux, nationaux, voire internationaux, plutôt recherche, plutôt économie, plutôt administration. Si on prend l'exemple de la santé-environnement, comment décliner tout cela pour répondre à l'enjeu d'acquérir une meilleure connaissance, par les chercheurs et par les professionnels de santé, de l'impact qu'a l'environnement sur la santé, de l'impact de la pollution de l'eau, de la pollution de l'air, du bruit, des vagues de chaleur, des pesticides sur la santé ? En l'espèce, nous constituons une première brique véritablement structurée autour de la donnée. Celle-ci devient de ce fait un enjeu pour rassembler tout un écosystème. On donne accès à cette donnée de qualité, en l'occurrence sous une forme catalogue, et ça devient un terreau pour faire travailler les acteurs et activer, stimuler l'utilisation effective de ces données pour avancer sur tous ces sujets. Nous avons alors toute une série de projets très concrets, de recherches

professionnelles qui vont pouvoir être imaginés. L'accès à la donnée donne aussi des idées de travaux, portées soit par des chercheurs, soit par des entreprises, soit par des administrations. Par exemple, comprendre pourquoi les fréquentations aux urgences de tel territoire sont plutôt comme ceci ou comme cela, pourquoi on a tel niveau de concentration de maladies des enfants à tel endroit… Faire des corrélations et influencer d'autant sur les politiques publiques, sans oublier qu'il y a une forme de prérequis extrêmement importante à savoir tout le travail fait sur la donnée. Dans le domaine de la transition écologique, les données sont souvent ouvertes depuis assez longtemps, parce qu'il existe toute une série de règles internationales qui font qu'on n'a pas attendu la décennie 2010, et notamment cette loi très importante, la loi Lemaire pour une République numérique, ou la loi Valter qui l'avait précédée un an plus tôt… On n'a pas attendu ces lois-là dans le domaine de l'environnement pour faire de l'*open data*. Même si la donnée est déjà ouverte massivement depuis assez longtemps, il reste un gros travail à faire pour qu'elle soit trouvable et réellement exploitable, notamment pour l'ensemble de ces projets-là. C'est un premier paramètre important. Comment agissons-nous là-dessus ? Il y a une dimension « politique de l'offre, politique de la demande ». Politique de la demande, c'est un peu ce que j'évoquais : dans le cadre de France 2030, sur 40 millions d'euros de projets, 20 millions d'euros ont été spécifiquement dédiés au recours effectif à l'intelligence artificielle par les collectivités territoriales pour leur transformation écologique. Là, on vient stimuler la demande d'un recours à l'intelligence artificielle et d'emblée, on se fixe pour objectif qu'elle soit frugale, ce qu'on évoquera un petit peu plus tard. Et à côté, il y a tout ce travail sur l'offre avec un label porté par l'État, « Greentech Innovation[5] ». On vient identifier des entreprises dont on estime que leur solution *data*, IA, va avoir un réel impact écologique. Dans le cadre d'un partenariat avec Hub France IA, une cartographie est en cours de l'ensemble des solutions privées, françaises, en matière d'intelligence artificielle, solutions dont on pense qu'elles ont une utilité réelle pour la transition écologique. C'est un vrai défi, parce qu'à la fois, il y a un double phénomène de *washing*, « blanchiment » en français, de *greenwashing* et d'*IA washing*. Il y a beaucoup de choses qui sont présentées comme étant de l'intelligence artificielle, qui ne le sont pas forcément, parce qu'il y a des jeux de marché qui font que c'est comme ça. Et parfois, des solutions sont présentées comme ayant un impact écologique, alors que ce n'est pas le cas. Nous faisons tout un travail pour lutter contre cela et mieux connaître cette offre pour mieux la favoriser, pour que les acteurs français – que ce soient des entreprises, des territoires, des citoyens, des associations – aient recours à ces solutions. Nous avons publié la semaine dernière au Salon des Maires de France la nouvelle version d'un livre blanc[6] exclusivement dédié à l'améliorations des relations entre ces mondes. Si l'on veut que tout ce qu'on discute s'incarne, à la fin il faut qu'il y ait des femmes et des hommes qui décident de collaborer, celles et ceux qui portent ces solutions liées à l'intelligence artificielle et celles et ceux qui en ont besoin pour leur transition écologique. Tout un travail a été fait pour mettre sur la table les difficultés et encourager les bonnes pratiques de collaboration, y compris des collaborations commerciales, avec la particularité propre au domaine de la transition écologique que l'on arrive très souvent dans le champ de la commande publique. Là, on retrouve un enjeu assez important dans le cadre de la commande publique et du droit européen de la commande publique : comment faire pour que les acheteurs publics soient plus facilement en confiance dans le fait d'aller travailler avec une entreprise qui sera peut-être un peu jeune, qui proposera une solution d'intelligence artificielle sur tel ou tel sujet, et en miroir, comment faire en sorte que les responsables de ces entreprises aient confiance dans la prise de risque que constitue le fait d'aller prendre un petit peu de temps, peut-être un peu plus, pour aller travailler avec des acteurs publics ? Ces enjeux ne sont pas forcément souvent mis en avant, comme du reste celui de la donnée, mais sont des prérequis extrêmement importants pour agir. Je rejoins complètement monsieur Le Goff sur la dimension formation et acculturation. Il y a tout un travail à faire pour sortir du premier cercle des personnes qui connaissent, qui comprennent. Depuis la sortie des outils d'intelligence artificielle générative comme ChatGPT, une grande partie de la population prend l'habitude de manipuler ces outils, malheureusement sans avoir conscience de leur impact environnemental. Là, nous organisons de façon assez classique des « cafés IA », qui ont été préconisés aussi par le rapport remis en début d'année au président de la République, et qui sont faits non pas pour celles et ceux qui travaillent

5 https://greentechinnovation.fr/appel-a-manifestation-dinteret/
6 https://greentechinnovation.fr/storage/V3-Livre-blanc-1.pdf

déjà dans l'intelligence artificielle, mais pour celles et ceux qui ont envie de comprendre comment l'IA pourrait s'intégrer dans leur mission. Là, c'est dans un cadre public. Et on en profite aussi pour pousser nos grands principes sur la meilleure façon de l'utiliser. On retrouve là toutes les questions transverses, et l'importance des cas d'usage. Par exemple, prenons le cas d'usage de la maîtrise de l'artificialisation des sols : l'IGN porte l'un des plus beaux projets dans le périmètre du pôle ministériel, une véritable pépite technologique française. Il s'agit d'un outil qui permet, grâce à l'intelligence artificielle et à l'apprentissage, d'automatiser le lien entre un pixel sur le sol et la compréhension de ce qu'il est. Tel ou tel volume de pixels, est-ce de la vigne, de la toiture, est-ce du parking, de la route, etc. ? L'IGN a conçu cet outil pour exploiter les images du territoire français qui sont prises en grande partie grâce à des LiDAR sur les avions (tous les ans, un tiers du territoire français est ainsi photographié). L'automatisation de la compréhension de l'affectation du sol permet ainsi de créer un outil qui a un potentiel très important pour mieux maîtriser l'artificialisation des sols. C'est extrêmement puissant pour passer ce message sur l'intérêt de l'intelligence artificielle. En objectivant la situation, on peut aussi lutter contre la façon dont ce sujet-là est parfois traité dans les médias, parfois aussi par des responsables politiques ou des influenceurs. Ça passe par de la formation, comme ça a été évoqué, mais aussi par beaucoup de sessions d'acculturation où il faut se mettre à niveau et s'adapter aux besoins métiers des agents publics. Ce sont les différents leviers qu'on utilise pour stimuler tout cela, en plus de notre cœur de métier de l'incubation de projets. Je tenais à insister sur ces dimensions-là qui ne sont pas forcément mises en avant et qui sont extrêmement importantes aussi.

Est-on capable de quantifier aujourd'hui l'impact écologique de l'IA et qui sont les acteurs qui l'évaluent ?

Thomas Le Goff : La question de l'empreinte environnementale de l'IA est un sujet absolument crucial. La réalité, c'est qu'on sait depuis un certain temps mesurer l'empreinte environnementale de l'intelligence artificielle. Les premières études quantifiant le besoin énergétique de l'entraînement des modèles ont déjà plus de cinq ans et elles se multiplient depuis. On sait donc mesurer les besoins énergétiques pour l'entraînement d'un modèle. Pour donner quelques chiffres, l'entraînement d'un grand modèle d'apprentissage automatique en traitement du langage, du même type que celui derrière ChatGPT, en 2019, c'était 500 tonnes de CO_2 juste pour l'entraînement. Depuis, d'autres études ont montré que la phase d'entraînement, ce n'est que 10 % des besoins globaux en énergie pour le fonctionnement d'un système d'IA. On connaît aussi l'empreinte en eau qui diffère beaucoup en fonction des types de *data centers*, de leur localisation, des normes environnementales qui leur sont applicables, etc. On n'a pas de connaissance globale de cette empreinte-là, mais on sait la quantifier de façon locale. Aujourd'hui, on sait qu'une conversation avec ChatGPT, par exemple, consomme l'équivalent d'environ un litre d'eau (on est à peu près à 40 millilitres par inférence, et on considère qu'une conversation, c'est entre 20 et 60 requêtes). Des outils sont développés à partir de ces données-là, par exemple, l'outil « Code Carbon ». Il existe aussi des outils de transparence pour donner des ordres de grandeur à des utilisateurs, comme ce qui a été développé par EcoLogits aussi. Hugging Face mène des travaux sur la réalisation d'un score carbone pour les modèles d'intelligence artificielle. J'ai parlé des besoins énergétiques et des besoins en eau. Il y a évidemment aussi les besoins en matériaux pour construire les puces électroniques, les *data centers*, les serveurs, etc., c'est-à-dire tout le *hardware* qu'il y a derrière l'immatériel du système d'IA qu'on utilise au quotidien. Là, on a moins de données. On peut citer quelques études comme celles de Kate Crawford aux États-Unis avec un livre intitulé *The Atlas of AI*, ou encore *Anatomy of AI* dans lequel la chercheuse a étudié toute la chaîne de valeur d'Amazon Echo, l'assistant vocal d'Amazon. Cette étude commence à dater un peu, mais essaie justement de retracer l'extraction des métaux rares, etc., qui sont nécessaires à la conception de ces systèmes-là. On note une multiplication des études qui essaient de quantifier tout cela, mais l'un des verrous est le manque d'études globales, parce que les données sont difficiles à agréger. Il faudrait analyser l'impact de tout et pouvoir agréger l'ensemble. On n'a pas une vision globale de cette empreinte-là, même si on sait la déterminer au cas par cas. La principale problématique, c'est que personne n'est d'accord sur la méthodologie, car on ne s'est jamais mis autour de la table pour se dire : « Que mesure-t-on, comment, quel est le *scope* d'émission que l'on va documenter et sur lequel on va faire du *reporting* ? Pour la consommation en eau, est-ce qu'on prend une moyenne, est-ce qu'on opte précisément pour le *data center* dans lequel un modèle a été

entraîné plutôt que celui sur lequel ce modèle fonctionne ? » Parfois, il n'est même pas évident de savoir dans quel *data center* la requête est envoyée pour obtenir une réponse… L'enjeu, aujourd'hui, porte vraiment sur le consensus, sur cette méthode d'évaluation. Sur les acteurs, vous les évoquiez, il y a des acteurs publics, mais je laisserai monsieur Cottinet développer cet aspect. Plus globalement sur les impacts environnementaux du numérique, l'Arcep et l'Ademe sont très actifs sur le sujet, le ministère en charge de la transition écologique évidemment, et on voit petit à petit une diffusion de ces problématiques-là auprès d'autres acteurs publics et des acteurs internationaux également.

L'Ecolab a récemment publié le *Référentiel pour une IA frugale*. Quelle est son ambition ?

Thomas Cottinet : J'ai évoqué tout à l'heure des cas d'usage. Une partie d'entre eux est déployée dans le dispositif des « démonstrateurs territoriaux d'intelligence artificielle frugale », dont sont partenaires le groupe Caisse des dépôts et consignations et la Banque des territoires dans le cadre de « France 2030 ». Quand nous avons imaginé ce dispositif il y a deux ans, d'emblée, nous nous sommes autocontraints pour que l'ensemble des projets émergent à partir d'intelligence artificielle *frugale*, même si la définition était assez floue à l'époque. Depuis, une communauté de plus de 900 membres a émergé, rassemblant des acteurs publics et privés de l'intelligence artificielle pour l'environnement. Et avec Guillaume Avrin, le coordinateur national de l'intelligence artificielle, nous est venue l'idée, à la fin de l'année 2023, d'interpeller l'ensemble des représentants de l'écosystème français, pour leur donner rendez-vous le 15 janvier 2024, pour leur proposer une ambition : doter la France, première étape de normalisation, d'un référentiel sur cette « intelligence artificielle frugale ». Ce qu'avaient démontré les deux premières années et l'ensemble de ces projets, c'est l'existence d'outils extrêmement intéressants. Au sein des équipes de recherche françaises, notamment le CNRS, il y a un groupe EcoInfo qui porte plusieurs outils, dont l'un, Green Algorithm, est absolument remarquable. Derrière cette ambition, il y a aussi, depuis le début, l'idée de stimuler les travaux de recherche pour accélérer. Plus d'une centaine d'acteurs publics et français sont venus au début de l'année et ont accepté de jouer le jeu. Au total, 150 acteurs ont travaillé pendant six mois. Pour le coup, on a mis les personnes autour de la table pour définir un premier référentiel et se fixer de premiers objectifs : définir ce qu'était l'intelligence artificielle frugale et commencer à travailler sur des méthodes, des pratiques pour garantir l'impact environnemental de l'intelligence artificielle. Ça a été beaucoup plus rapide que prévu et dès le mois de juin, la France a publié son *Référentiel général pour l'intelligence artificielle frugale*, une collaboration entre l'Ecolab du Commissariat général au développement durable du ministère et l'AFNOR. Et ça a été une première mondiale, ce dont on n'avait pas forcément conscience au départ. La France ainsi a été le premier pays au monde à faire cette publication officielle sur ce sujet très complexe. Cette réussite nous a permis très rapidement de proposer que les institutions européennes s'en saisissent *via* le Comité européen de normalisation et son Joint Technical Committee 21. Et là, l'accueil a été très favorable. Il faut rappeler qu'en arrière-plan de tout cela, il y a eu la validation d'un règlement européen majeur, l'« AI Act », qui comporte certaines dimensions environnementales, mais qui, disons-le, n'est pas très ambitieux dans ce domaine-là. D'où le souci des acteurs de compenser cela. Depuis l'été, depuis le 10 juin précisément, tout un travail a été enclenché pour doter l'Europe d'une couche de normalisation sur ce sujet-là. Et comme cela se passait plutôt bien, assez tôt est venue l'ambition de passer cela à l'échelle internationale, ce qui s'est fait avec plusieurs organisations internationales : l'OCDE, l'Organisation internationale des télécoms, et surtout l'Unesco. Là aussi, nous avons reçu un accueil assez favorable. Au final, une ambition : produire une feuille de route mondiale de l'impact environnemental de l'intelligence artificielle. La France accueille en février 2025 le sommet mondial pour l'IA. « L'intelligence artificielle durable » va être un des livrables de ce sommet et constitue un jalon. L'ensemble des groupes de travail ont ce sommet en ligne de mire. Et tout est fait pour que nous soyons capables, en février 2025, de proposer des outils très concrets qui auront répondu à une partie des questions que pose Thomas Le Goff, parce que, pour le coup, on met les sujets sur la table et on essaie de combler ces vides, qui existent aussi bien dans la littérature scientifique que dans les pratiques. Ces manques sont parfois virtuels : les gens ne se rendent pas forcément compte qu'à tel endroit dans le monde ou dans telle filière, en réalité, il y a déjà telle ou telle initiative. Tout un travail est fait pour rassembler les initiatives et pour créer une nouvelle façon d'inciter à une meilleure maîtrise de l'impact

environnemental de l'IA. Cela signifie aussi parfois suggérer de ne *pas* recourir à l'intelligence artificielle. Et quand il y a recours à l'intelligence artificielle, de ne pas forcément recourir aux modèles qui sont les plus énergivores, émetteurs de CO_2, consommateurs d'eau ou de métaux rares. Nous sommes en mesure de faire connaître toute une série de solutions d'intelligence artificielle autres que l'IA générative et qui restent extrêmement intéressantes. Sur le non-recours (carrément) ou sur le recours à des solutions d'intelligence artificielle adaptées, nous avons besoin de diffuser les bonnes pratiques, d'identifier des outils comme des façons de faire et des modes d'évaluation. C'est tout cela qui est en cours. Nous avons l'ambition de nous saisir de cette opportunité qu'est le sommet mondial pour l'intelligence artificielle de février 2025 sous le patronage du président de la République pour accélérer des annonces d'investissement dans des *data centers* vertueux, ce qui implique d'être capable de définir ce qu'est un *data center* vertueux. Cela stimule les travaux en cours sur ce sujet-là. On est plutôt dans un bon *momentum*, même si la situation économique peut être compliquée dans certains pays. On observe une amplification du sujet tirée par le sommet et par des stratégies économiques industrielles. On a beaucoup mis en scène l'opposition qu'il peut y avoir entre la régulation et la compétitivité ; il n'empêche que de grands acteurs économiques industriels sont en train de reconnaître que cela peut être une opportunité, un avantage concurrentiel que de faire partie des premiers acteurs économiques qui seront en capacité d'attester que leurs solutions d'intelligence artificielle sont vertueuses. Ceci dit, on part de très loin. Google, qui était il y a quelques années une des entreprises les plus intéressantes en termes de trajectoire nette zéro, a un peu déraillé à cause de l'intelligence artificielle générative, qui représente une hausse de 40 % d'émissions de tonnes de CO_2 entre 2019 et 2023. C'est beaucoup. Chez Microsoft, les ordres de grandeur sont à peu près les mêmes. Sans être naïfs, quelque chose est en train de se passer avec un alignement. Chacun joue son rôle. La recherche scientifique est fortement impliquée et accroît la production de connaissances pour mieux évaluer l'impact et faire des préconisations de solutions techniques qui contribueront à un impact plus faible. La communauté des sciences humaines et sociales est également présente, avec notamment les juristes, la *legal tech*, pour produire des référentiels innovants. Les industriels, petits et grands, sont aussi autour de la table, font des propositions, intègrent cela, en ligne soit avec leur stratégie RSE, soit avec leur stratégie commerciale. Les pouvoirs publics, les administrations, portent tout cela, le soutiennent. En France, le ministère de la Transition écologique a été chargé de coordonner tout cela pour le compte de l'ensemble du gouvernement. Le fait que le sommet mondial pour l'IA aura lieu en France nous place en ce moment en coordinateur au niveau international. On est très exposé sur ces sujets-là, et on a bien conscience de cette complexité et de tous ces défis à relever que rappelait Thomas Le Goff. On est plutôt dans un moment où ça bouge. Des enjeux économiques peuvent être contraires, mais il n'empêche qu'une très forte accélération s'est produite cette année, avec des espoirs gigantesques pour qu'elle continue au moins jusqu'au sommet mondial pour l'IA, avec, derrière, un effet étincelle. Nous avons l'objectif que les livrables qui seront validés au niveau international par l'ensemble des parties prenantes se mettent en œuvre et se déploient. Il s'agit aussi pour nous de mettre en place une coalition internationale de l'IA durable qui sera outillée pour faire advenir et prospérer l'ensemble des objectifs qui seront validés en matière d'intelligence artificielle pour l'environnement lors de ce sommet.

Thomas Le Goff : Je voudrais juste ajouter un mot sur le rôle de la régulation. Concernant l'opposition entre la régulation et l'innovation/la compétitivité, qui est un sujet important, je pense qu'il faut rappeler que toute innovation n'est pas signe de progrès. Un des rôles de la régulation, c'est de s'assurer que l'innovation se fasse dans une trajectoire souhaitable d'un point de vue social. Dans la recherche en droit et en économie, cela se traduisait classiquement par le fait que la régulation était considérée comme un outil pour corriger des imperfections de marché. Aujourd'hui, sur le marché de l'IA, on assiste au phénomène suivant : des externalités négatives commencent à prendre une place peut-être un peu trop importante, avec justement ces dérives sur les objectifs de neutralité carbone, etc. Notre rôle est de réfléchir aux leviers de régulation qui peuvent être mis en œuvre pour corriger cette trajectoire. Cela impose d'abord de se poser la question suivante : les acteurs au sens large (pas uniquement les fournisseurs d'IA, mais également les acteurs publics qui vont essayer d'inciter certains comportements), et le marché sont-ils capables de se corriger et d'adopter une trajectoire durable sans la contrainte ? Si la réponse est non, alors une réglementation plus contraignante est nécessaire. Vous avez évoqué le règlement européen sur l'IA qui était une occasion d'avoir des obligations contraignantes sur cette

question-là, et notamment une réglementation qui vise à impliquer des valeurs européennes dans le développement de la technologie *via* des mécanismes *ex ante*, c'est-à-dire des obligations de conformité dans le processus de conception. C'est le parfait endroit pour prendre en compte des questions de durabilité. Même si des dispositions relatives à l'environnement étaient apparues au cours du processus législatif, ce qui en sort à la fin est assez décevant parce qu'on s'en tient à des obligations de documentation, ce qui impose aux fournisseurs d'IA de documenter, donc tracer un certain nombre d'éléments relatifs aux besoins en ressources informatiques, et uniquement dans la phase d'entraînement. Tout un pan de l'impact environnemental qu'on a évoqué sur la consommation en eau, en matériaux, etc. n'est pas du tout couvert. De surcroît, ces obligations ne s'imposent qu'à certains systèmes d'IA, notamment les systèmes d'IA à haut risque, les IA utilisées dans le secteur médical, en tant que composants de sécurité dans des infrastructures critiques. Ce ne sont pas les systèmes d'IA les plus gros, mais plutôt des systèmes relativement spécialisés. Est-ce qu'on ne se trompe pas de cible ? En revanche, ces obligations s'appliquent – c'est positif, il faut le souligner – aux grands modèles de langues, par exemple, *via* les dispositions relatives aux systèmes d'IA à usage général. Ce sont des systèmes qui reposent sur un apprentissage à partir d'importantes quantités de données applicables à de très nombreuses finalités. C'est positif, même si le contenu de l'obligation de documentation est relativement décevant. Ce qui manque également, c'est une obligation de transparence. Il n'y a pas de contrainte pour le fournisseur ou le « déployeur » d'un système d'IA d'informer l'utilisateur sur ses impacts environnementaux, alors qu'en termes de régulation, ce serait très intéressant pour potentiellement orienter la demande et se reposer sur les choix des consommateurs qui, ayant une information complète, pourraient réaliser des choix éclairés et peut-être prendre en compte des critères environnementaux. La question de la transparence est essentielle. Le développement de tous ces standards peut être le présage de réglementations futures plus contraignantes dans l'hypothèse où l'adoption volontaire serait insuffisante. C'est un point très important de la question de la régulation qui est en train de se jouer aujourd'hui. Nous allons devoir suivre la façon dont ces travaux impactent la production et la publication de normes techniques relatives à l'application du règlement européen sur l'IA, parce que toutes ces initiatives peuvent influer sur ce processus normatif. Une forme de porosité serait très vertueuse. Par ailleurs, nous devrons surveiller l'adoption volontaire de ces standards non contraignants qui sont établis aux niveaux français et européen, et – espérons-le – mondial, pour évaluer leur effectivité. Voilà en quoi consistera selon moi le travail de la recherche dans les années qui viennent.

Quelle est la place du citoyen-usager dans cette dynamique d'une IA frugale et comment favoriser la transparence sur ces questions ?

Thomas Cottinet : Les citoyens sont une cible et au travers de certaines associations, ils participent à la construction de ces référentiels. Dans le cadre des travaux que nous avons enclenchés en essayant d'utiliser le sommet mondial pour l'IA comme accélérateur, nous menons un travail sur un observatoire de l'intelligence artificielle durable, frugale, qui mette en lumière l'impact environnemental de l'intelligence artificielle. On retrouve le sujet de la donnée, de la connaissance, une forme de mise en abyme. On a aussi des outils qui permettent de partager plus largement la connaissance et les données auprès des citoyens. Sur cet axe, les écosystèmes sont un peu les mêmes que ceux qui travaillent sur l'ensemble des outils non contraignants, normes et standards, avec à chaque fois une configuration « public/privé » que je trouve extrêmement intéressante. On se concentre en priorité sur des logiques institutionnelles très puissantes, comme celle de l'adoption d'un règlement européen, c'est tout à fait normal, mais nous ne voulons pas pour autant laisser de côté d'autres démarches descendantes et remontantes. Beaucoup d'initiatives voient le jour en ce moment. Il y a une rétroaction entre les initiatives qui veulent être prises par des citoyens, des entreprises, des chercheurs et la façon dont c'est régulé. Par exemple, la semaine du 4 décembre 2024, en lien avec une entreprise, Hugging Face, et avec Data for Good[7], nous lançons le « Défi de l'intelligence artificielle frugale ». On retrouve des acteurs privés, publics, de l'argent privé, public aussi, pour stimuler tout cela. À

7 https://dataforgood.fr/

chaque fois, les outils et initiatives qui émergent vont initier de nouvelles façons de réguler et d'encadrer les différentes parties-prenantes. Tout cela s'auto-alimente avec les citoyennes et les citoyens qui sont au cœur de la démarche. On fait un focus sur l'impact environnemental de l'IA. Les enjeux éthiques sont bien pris en compte par le règlement européen sur l'IA : ils concernent l'explicabilité, les biais, les risques d'intrusion, de surveillance, les freins dans l'accès à des services publics ou des services privés. La société civile participe pleinement à la controverse. S'agissant de l'impact environnemental, l'Observatoire, les politiques publiques et leur communication, la façon dont les filières s'en saisissent et cette dynamique autour de ces référentiels qui certes n'ont pas le même niveau de contrainte que pourrait avoir une loi européenne, mais qui constituent des avancées, tout cela est propice à l'information de l'ensemble de ces écosystèmes. À titre personnel, en tant que citoyen, puisque vous me posez la question « *Quid* des citoyens ? », je regrette que ces débats soient peu politisés et peu présents dans le débat public. Le Parlement, les grandes échéances publiques, pourraient mettre bien plus en avant ces sujets-là. Cela contribuerait à les rendre accessibles à l'ensemble des citoyens, et à faire comprendre pourquoi des questions se posent. Les rares fois où cela arrive dans le débat public, c'est de façon négative, parce qu'il y a une crainte, un scandale. On se rend compte de l'existence de tel ou tel biais dans telle ou telle façon d'accorder ou pas une allocation, ou de déclencher un contrôle sur des individus. La jambe négative-défensive est bien là, qui alerte et fait peur, mais la jambe offensive-positive n'est pas très présente. Les débats d'experts prévalent, comme ceux que nous sommes en train d'en avoir, alors qu'il y a vraiment besoin de faire déboucher le thème dans le débat public, certes au prix d'un gros travail d'acculturation et de vulgarisation. À ce titre-là, je trouve que les initiatives que nous prenons avec les agents publics qui y contribuent, agents qui sont aussi des citoyennes et des citoyens, cet investissement fait pour outiller, acculturer, expliquer, faire en sorte que l'ensemble des agents publics qui le souhaitent accèdent à une forme de connaissance, de compréhension des enjeux qui se posent au travers de l'intelligence artificielle et aient l'occasion, à bon escient, de l'intégrer dans leur pratique professionnelle, cela va apporter une pierre à notre édifice, mais ce n'est pas suffisant. Pour ma part, je plaide pour une *politisation*, au bon sens du terme, beaucoup plus forte de ces sujets, et que soit mis sur la table ce qui se joue, quand on décide de recourir à telle ou telle solution, par rapport à des libertés publiques, par rapport à un impact positif. On s'est dit tout à l'heure que l'IA était une des solutions potentielles pour relever les défis de l'urgence climatique, sans négliger certains risques éthiques, sociaux, environnementaux. On parle beaucoup en ce moment de politiques publiques sous l'angle de la transformation publique, des enjeux qui pèsent sur les finances publiques et des économies à envisager. Si l'intelligence artificielle, ce qui est déjà le cas, est utilisée comme une partie de la solution, il faut être en capacité de le faire en connaissance de cause et en étant complètement transparent vis-à-vis des citoyennes et des citoyens qui vont bénéficier de ces services publics parfois déjà mis en place à partir d'intelligence artificielle. Comment est-on totalement transparent ? Comment donne-t-on les clés de la connaissance et de la compréhension sur ces sujets-là ? Je trouve que c'est l'un des grands défis à relever. Pour le coup, cela peut paraître superficiel d'organiser beaucoup d'événements, de conférences, mais je trouve au contraire que c'est extrêmement intéressant, car cela participe à tout ce travail. J'espère que le sommet mondial pour l'IA qui aura lieu en France sera une belle occasion de diffuser davantage, d'acculturer et de partager la connaissance sur l'intérêt de l'intelligence artificielle dans toute une série de matières, les risques qui peuvent y être associés, la façon dont il faut l'encadrer pour pouvoir l'utiliser à bon escient et en connaissance de cause, et enfin les enjeux environnementaux, éthiques et sociaux qui y sont associés.

Thomas Le Goff : Pour dire un mot sur la relation entre la régulation, le citoyen et l'usage, je crois beaucoup dans le rôle de la société civile *via* son influence sur les choix technologiques et sur l'impact que peuvent avoir l'usage d'une technologie et les choix qui sont réalisés par les usagers sur une technologie. Prenons un exemple. Des entreprises m'ont dit qu'elles ont déployé des outils d'IA générative en interne pour augmenter la productivité et que des collectifs de salariés se sont opposés à l'utilisation de cet outil en raison de considérations écologiques. Finalement, je crois que la régulation est un outil pour donner le pouvoir aux usagers. J'ai évoqué l'idée de la transparence. Cette prise de pouvoir par les usagers n'est possible qu'à partir du moment où ils ont une information complète, où ils sont sensibilisés. Une fois que les personnes sont sensibilisées, la réalité, c'est que cela a de vrais impacts sur leur usage de la technologie. Cet effet est plutôt bien documenté en termes sociologiques si l'on fait une analogie avec d'autres secteurs d'activités polluantes par le passé. Cela doit être selon moi un vrai objectif, à

la fois de politique publique, mais aussi dans toute potentielle initiative de régulation, que de mettre les citoyens au centre et de leur donner le pouvoir d'action, en leur donnant aussi la possibilité de tenir pour responsables les fournisseurs d'IA qui ne seraient pas vertueux. Cela passe par la mise en place d'obligations contraignantes et de mécanismes de plainte, comme c'est le cas dans l'« AI Act », même si la partie environnementale est assez décevante. Ou encore, par les obligations en termes de responsabilité sociale et environnementale faites aux entreprises de diffuser de l'information de façon transparente sur leurs impacts environnementaux et qui, demain, *via* le devoir de vigilance en France, et demain en Europe, permettront aussi à des collectifs, *via* des actions de groupe, de tenir pour responsables des personnes, des entités qui soit auraient pris des engagements beaucoup trop faibles, soit ne mettraient pas les mesures suffisantes au regard de leurs objectifs. Je pense que cette évolution de la régulation aura partie liée à cette dynamique, qui me semble vraiment essentielle, consistant à donner plus de pouvoir aux usagers.

Quels sont les prochains rendez-vous ?

Thomas Le Goff : Je voudrais ajouter que, dans le cadre du sommet IA, des journées scientifiques seront organisées par l'Institut polytechnique de Paris les 6 et 7 février 2025 sur la thématique de la place de l'IA dans la science et dans la société. De nombreuses tables-rondes seront consacrées aux questions d'éthique, de biais, de maths, aux sciences dures et aux sciences sociales également. Il y aura un *workshop* sur la thématique « IA et environnement » co-organisé avec Juliette Fropier de l'Ecolab, pour faire entrer cette dimension dans le milieu scientifique puisque la réalité, c'est que jusqu'à très récemment, les organisateurs trouvaient aisément des personnes pour assurer tous les autres *workshops*, mais qu'il n'y avait pas assez de chercheurs travaillant sur ces aspects. C'est important de le diffuser à cette échelle.

Thomas Cottinet : La semaine du 4 décembre se tiendra une importante réunion de l'agence internationale de l'énergie sur ce sujet. Le lancement du « Défi de l'IA frugale » est imminent. Nous préparons aussi, en prévision du sommet mondial de l'IA le 11 février, un *side event* exclusivement dédié à l'ensemble des livrables du sommet qui traitent d'environnement. Il y a une dynamique. Ces échéances, ces objectifs nous permettent d'accélérer tous ensemble sur ces sujets.

Études

Expliquer ou justifier : comment s'outiller pour permettre un déploiement des systèmes de décision algorithmiques de confiance ?

Clément Henin

Cet article explore les enjeux liés à l'utilisation croissante des systèmes de décision algorithmique (SDA) dans les administrations et entreprises. S'appuyant sur des concepts comme la transparence, la justification et la contestabilité, il met en lumière les risques d'opacité et de décisions injustifiables qui menacent la légitimité et la confiance des citoyens. Une expérimentation auprès de la CNIL est détaillée, testant deux outils – IBEX pour l'explication et Algocate pour la justification – et révélant leur potentiel à améliorer l'audit des algorithmes. L'article propose des pistes pour renforcer l'éthique et la supervision des SDA, essentiels à leur déploiement responsable dans les services publics.

Les systèmes de décision algorithmiques (SDA) sont de plus en plus utilisés dans divers domaines du secteur privé comme du public. Certains SDA reposent sur l'apprentissage automatique, alors que d'autres sont basés sur des règles prédéfinies. Certains prennent des décisions de manière entièrement automatisée, tandis que d'autres se contentent de fournir une aide aux décideurs humains. De plus, certains de ces systèmes sont destinés aux professionnels, là où d'autres visent davantage le grand public. Leur utilisation n'est pas anecdotique et de nombreuses entreprises et administrations en exploitent déjà. Des SDA sont ainsi employés dans des secteurs variés comme le transport (véhicules autonomes), l'éducation (appariement élèves/formations), la sécurité (surveillance vidéo), les ressources humaines (filtre de recrutement), la médecine (diagnostics, appariements [Henin, 2021]), etc. Les SDA pourraient améliorer les processus décisionnels, en les rendant notamment plus rapides, reproductibles et économiques contribuant ainsi à accroître la qualité des services rendus. Cette amélioration repose cependant sur le fait que les algorithmes eux-mêmes soient efficaces et respectent les principes de transparence et de responsabilité. Confier à des SDA le pouvoir de prendre ou d'influencer des décisions potentiellement sensibles soulève des questions juridiques, éthiques, politiques et techniques. Ignorer ces questions pourrait mener à l'adoption de SDA dont les conséquences seraient préjudiciables pour les individus (discrimination, perte d'autonomie, etc.), pour l'économie (pratiques déloyales, accès limité aux marchés, etc.) et pour la société dans son ensemble (manipulation, menace à la démocratie, perte de confiance dans les institutions, etc.). Pour aborder efficacement ces questions, il paraît donc essentiel de centrer les réflexions sur la finalité des SDA, c'est-à-dire prendre une décision, plutôt que sur une technique particulière d'algorithme traditionnel ou d'intelligence artificielle (IA), bien que l'essor de cette dernière constitue un élément de contexte qui doit être pris en compte. Bien souvent, l'usage d'une technologie particulière (IA générative par exemple) obère, dès la phase de conception, certaines questions centrales de légitimité du système ou de justification des décisions prises par les algorithmes. Plus que les

techniques mises en œuvre, ce sont pourtant ces questions qui peuvent altérer la confiance des citoyens, voire rendre les systèmes inacceptables. Dans un système public de vidéosurveillance par reconnaissance faciale, ce n'est pas l'usage de l'apprentissage profond qui soulève des problèmes éthiques, mais sa finalité d'identification des individus dans un espace public.

Les exigences de transparence, d'équité et de responsabilité sont souvent mises en avant comme des solutions aux risques posés par les SDA. Cependant, ces notions sont fréquemment énoncées de manière vague et leur mise en œuvre concrète est très rare. Par exemple, des termes comme « transparence », « explication » et « justification » sont souvent mentionnés, mais l'absence de définitions précises conduit à des interprétations divergentes voire à des malentendus. Sur le plan technique, des développements récents ont permis de réduire les biais, d'améliorer la fiabilité et de générer des explications censées rendre le fonctionnement des SDA plus compréhensible. Toutefois, ces outils, notamment développés dans le champ de recherche *eXplainable Artificial Intelligence* (XAI), visent pour la plupart à répondre aux besoins de ceux qui en sont à l'origine : les chercheurs et les spécialistes en IA (Miller, Tim *et al.*, 2017). Les explications d'algorithme, par exemple, n'ont pas toujours les caractéristiques adaptées aux besoins des utilisateurs profanes, ces méthodes d'explications étant d'ailleurs rarement évaluées par des utilisateurs humains et encore moins par des professionnels non spécialistes de l'IA.

L'objectif de cet article est de mieux définir certains concepts essentiels (voir encadré ci-dessous) et de montrer comment, appliqués aux SDA, ils permettent d'en lever (pour partie) l'opacité. Ces concepts sont ensuite mobilisés pour rendre compte et analyser une étude expérimentale réalisée auprès d'agents de la fonction publique, afin de tester deux outils visant à mieux encadrer l'usage des SDA.

Concepts mobilisés autour des SDA : clarifications terminologiques

Les récentes avancées de l'IA et son utilisation pour les SDA ont donné lieu à d'abondantes publications dans des disciplines variées (informatique, droit, sciences sociales, sciences politiques, etc.). Les mêmes termes sont parfois utilisés dans des sens différents selon les disciplines, voire au sein d'une même communauté scientifique. Il est donc utile de fournir quelques clarifications terminologiques.

Système de décision algorithmique (SDA) : nous utilisons l'expression « système de décision algorithmique » plutôt qu'« algorithme » pour rappeler que ces systèmes ne se limitent pas à des algorithmes ou des logiciels. Ils doivent être appréhendés dans un cadre général intégrant leurs paramètres, leurs données d'apprentissage ainsi que l'organisation humaine (entreprise, administration, cadre juridique, etc.) et technique (serveur, logiciel, etc.) permettant leur fonctionnement.

Transparence : elle consiste à exposer (*via* la publication de codes ou de documents) le fonctionnement des SDA. Cette mesure organisationnelle ou réglementaire, impliquant les personnes responsables du SDA, permet de limiter l'opacité lorsque cette dernière est intentionnelle. Bien que souhaitable dans de nombreuses situations, elle est parfois insuffisante pour répondre à tous les enjeux notamment lorsque le fonctionnement du système n'est pas facilement intelligible.

Explication, explicable, explicabilité : une explication vise à permettre à une personne (concepteur, utilisateur, personne affectée, etc.) de comprendre le fonctionnement global du SDA ou une décision particulière. Par exemple, l'explication du refus d'une demande d'admission dans un établissement universitaire pourrait se fonder sur certaines notes (jugées insuffisantes) du dossier du candidat. Les explications renseignent soit sur la logique du SDA dans son ensemble (explication globale), soit sur la logique ayant conduit à une décision particulière (explication locale).

Justification, justifiable, justifiabilité : le but d'une justification est de convaincre qu'une décision est bonne (ou adéquate). Par exemple, une justification du refus d'admission dans une université pourrait s'appuyer sur le fait que les dossiers avec des notes faibles ont une probabilité d'échec élevée en fin de première année. Une autre forme de justification pourrait reposer sur le principe d'équité de traitement au regard d'autres candidats ayant obtenu de meilleures notes.

Contestation, contestable, contestabilité : l'objectif d'une contestation est de convaincre qu'une décision est mauvaise (ou inadéquate). La contestation d'une décision peut par exemple s'appuyer sur le fait que le candidat a obtenu des résultats exceptionnels dans d'autres matières ou que ses notes insuffisantes ne concernent qu'une seule année de sa scolarité.

Responsable, redevable[1] : selon Reuben Binns, « une partie A est redevable envers une partie B à l'égard de sa conduite C, si A est dans l'obligation de fournir à B une justification de C et peut être sanctionnée si B juge que sa justification est inadéquate »[2] (Binns, 2018, p. 544). Nous adoptons cette définition et considérons que les justifications sont un élément essentiel de la redevabilité. Comme les justifications peuvent être jugées « inadéquates », la contestabilité est également une condition nécessaire à la responsabilité.

Légitime, légitimité : de nombreuses définitions de la légitimité ont été proposées par les politologues, juristes et philosophes. Nous retenons la définition de Mark Suchman, suffisamment générale pour être applicable. Suchman définit la légitimité comme « une perception générale ou une hypothèse selon laquelle les actions d'une entité sont souhaitables, appropriées ou adéquates d'après un ensemble, construit par le corps social, de normes, de valeurs, de croyances et de définitions » (Suchman, 1995)[3].

Une approche fondée sur la transparence et l'explication insuffisante pour répondre à l'ensemble des enjeux

Dans cet article, nous faisons une distinction nette entre explication et justification, deux termes souvent confondus dans le champ de l'XAI. Les explications sont descriptives et intrinsèques, se limitant à faciliter la compréhension du fonctionnement du SDA, tandis que les justifications sont normatives et extrinsèques, reposant sur des références indépendantes du SDA permettant d'évaluer le bien-fondé des décisions. Afin d'affirmer qu'un résultat est bon (ou adéquat), il est nécessaire de se référer à une « norme » extérieure. Nous utiliserons le terme « norme » sans connotation juridique. Une norme peut en effet être une exigence légale ou réglementaire, mais aussi un principe éthique ou un objectif propre à l'organisation qui déploie le SDA. Dans l'exemple de la demande d'admission, la première justification est basée sur un objectif de l'organisation (minimisation des risques d'échec de ses étudiants) tandis que la seconde s'appuie sur un principe d'équité de traitement (pouvant lui-même être d'ordre législatif).

Bien que les explications et les justifications puissent contribuer à réduire l'opacité des SDA, elles poursuivent des objectifs fondamentalement distincts. En effet, un utilisateur peut comprendre la logique menant à un résultat particulier sans pour autant accepter que ce résultat soit bon. Et à l'inverse, il peut contester un résultat (convaincu qu'il est mauvais) sans comprendre la logique employée par le système. Certains juristes insistent sur l'importance de cette distinction et soutiennent que les décisions d'un SDA doivent pouvoir être justifiées indépendamment du fonctionnement technique du système[4]. Par exemple, il serait peu envisageable qu'une personne concernée par une décision administrative rendue par un réseau de neurones (bien que cela reste hypothétique à ce jour) reçoive des informations détaillant les liens entre les entrées et les sorties de cet algorithme complexe comme préalable à un possible recours. Les justifications et les contestations émergent difficilement dans le domaine de recherche en XAI dominé par des chercheurs en IA dont l'objectif principal est de concevoir des explications adaptées à leurs propres besoins, comme l'amélioration des algorithmes, dans des formes

1 Traduction en français du terme *accountable*.
2 Citation traduite par l'auteur.
3 Citation traduite par l'auteur.
4 La juriste Mireille Hildebrandt choisit l'exemple particulièrement éclairant des cours de justice : « Lorsqu'un tribunal statue sur une affaire, il ne peut pas justifier sa décision en exposant les heuristiques du ou des juges impliqués, comme leurs préférences politiques, ce qu'ils ont mangé au petit déjeuner ou la manière dont ils ont préparé l'affaire [...] la loi exige qu'ils motivent leurs décisions en se référant à un ensemble de raisons juridiques disponibles. » (Hildebrandt, 2019) Citation traduite par l'auteur.

et des niveaux de détail non pertinents pour la majorité des utilisateurs.

L'utilisation des données croît depuis plusieurs décennies, portée par des facteurs sociétaux (Porter, 1996) et des avancées techniques (capacités de stockage et de calcul). Cependant, les universitaires, les ONG et la société civile expriment des inquiétudes quant à leur impact, en particulier dans le cadre des SDA. Ce débat est souvent obscurci par des confusions et l'usage d'arguments de natures très différentes. Par exemple, on confond parfois la question de la légitimité des objectifs du système lui-même (par exemple surveiller l'espace public) avec celle, plus technique, des choix de mise en œuvre (l'utilisation de caméras augmentées avec des logiciels). Cette partie résume les principales préoccupations soulevées par le déploiement des SDA, avant de justifier les exigences auxquelles ces systèmes devraient répondre.

La justifiabilité et la contestabilité : des conditions nécessaires à la légitimité des décisions algorithmiques

L'usage croissant des SDA dans nos sociétés a conduit certains auteurs à évoquer le risque d'une « algocratie » ou d'une « gouvernementalité algorithmique »[5]. Selon Antoinette Rouvroy, « cette "gouvernementalité algorithmique" est caractérisée par des normes autoappliquées, implicites et statistiquement établies, émanant, en temps réel, de la réalité numérisée ; elle contraste avec la "gouvernementalité politique" et ses lois imparfaitement appliquées, explicites, délibérées d'une longue délibération politique »[6] (Rouvroy, 2013). Ainsi, l'un des principaux problèmes posés par les SDA est le caractère indiscutable, ou « non contestable », de décisions « autojustifiées » par l'algorithme lui-même sans avoir à en référer à un principe supérieur. Une autre composante clé de la légitimité de ces décisions est la redevabilité, c'est-à-dire l'obligation pour celui qui applique les décisions, d'en rendre compte.

Le droit positif reconnaît déjà la possibilité de contester les résultats des SDA. Cependant, les textes actuels se limitent en fait à un droit à des explications et à une intervention humaine. Par exemple, selon le considérant 71 du RGPD, lorsqu'un traitement automatisé est utilisé pour prendre une décision concernant une personne, celle-ci a « le droit d'obtenir une intervention humaine, d'exprimer son point de vue, d'obtenir une explication quant à la décision prise à l'issue de ce type d'évaluation et de contester la décision ». Les lignes directrices en matière d'éthique pour une IA digne de confiance publiées par le groupe d'experts de la Commission européenne (HLEG-AI) mentionnent également l'explicabilité comme un des quatre grands principes pour une IA digne de confiance (Commission européenne, Direction générale des réseaux de communication, du contenu et des technologies, 2019). Il est intéressant de noter que dans le RGPD aussi bien que dans les propositions du HLEG-AI, l'exigence d'explicabilité est suivie (et implicitement motivée) par la nécessité de permettre aux personnes concernées de contester les décisions. Cependant, comme nous l'avons montré précédemment, disposer d'explications ne suffit pas pour pouvoir justifier ou contester des décisions. Mireille Hildebrandt abonde en ce sens en précisant qu'« il ne faut pas que le discours de l'explicabilité fasse obstacle à la question de savoir si une décision est juridiquement justifiée, ce qui exige un type particulier de raisons juridiques. L'explication en soi n'implique pas la justification, et la justification n'exige pas toujours une explication de la logique sous-jacente du système décisionnel. »[7]

Contrairement aux explications, les justifications s'appuient sur des normes, c'est-à-dire des références externes du système algorithmique. Elles permettent ainsi d'éviter le risque de « l'autoproduction » de normes, émanant du système lui-même, sans référence ou contrôle extérieur. Ces normes sont variées, avec des sources de légitimité différentes et des formes multiples.

La contestabilité, garante d'une collaboration équilibrée et mutuellement bénéfique entre les SDA et les humains

Outre leur rôle essentiel pour assurer la légitimité de ces dispositifs, la justifiabilité et la contestabilité peuvent transformer radicalement la relation entre les SDA et leurs utilisateurs. Comme l'affirment Daniel N. Kluttz et ses coauteurs, « la contestabilité favorise l'engagement plutôt que la passivité, le questionnement plutôt que l'acquiescement. […] La contestabilité peut promouvoir un engagement

5 Ces termes sont empruntés respectivement aux juristes John Danaher et Antoinette Rouvroy.
6 Citation traduite par l'auteur.
7 Citation traduite par l'auteur.

critique, enrichissant et responsable entre les utilisateurs et les algorithmes. » (Kluttz *et al.*, 2020)[8] Ainsi, la contestabilité contribue à préserver l'autonomie du décideur humain. En effet, bien que ce dernier ne soit pas toujours tenu de suivre le SDA, son autonomie est en pratique limitée s'il ne dispose pas des moyens nécessaires ni d'outils efficaces pour remettre en question ces recommandations.

Dans un contexte où les administrations sont incitées à utiliser des SDA, notamment ceux s'appuyant sur de l'IA, pour mettre en œuvre une partie de leurs missions (ciblage des contrôles, *matching*, processus d'optimisation, etc.), la possibilité de contester ces systèmes et l'explicitation précise des objectifs et textes juridiques encadrant ces processus automatisés sont une condition indispensable d'un déploiement légitime. L'action publique intervient de façon exclusive dans de nombreux domaines et rend donc inévitable pour les citoyens certains SDA qu'elle met en œuvre. Les exigences des algorithmes dans le secteur public devraient donc être plus importantes encore que dans le secteur marchand. Développer des IA de confiance ne passe pas uniquement par des moyens techniques et des algorithmes fiables et précis ; cela passe en également par la possibilité laissée aux humains (professionnels, experts ou profanes) de superviser ces systèmes, de vérifier, à tout instant de leur cycle de vie, l'adéquation des résultats avec les objectifs initialement fixés et d'intervenir sans avoir à maîtriser ni à manipuler des concepts techniques. La confiance des citoyens va dans les systèmes capables de prouver qu'ils sont en phase avec les objectifs et les règles qui leur sont assignés, indépendamment de leur fonctionnement technique que seuls les experts ont vocation à comprendre.

Pour mettre en œuvre de tels systèmes en respectant les principes éthiques qui devraient guider leur action, les services de l'État auraient vocation à mieux appréhender ces SDA au regard de leurs conséquences sur les citoyens et indépendamment des techniques utilisées pour les déployer. Au-delà d'une nécessaire acculturation des cadres et décideurs de l'administration, cela passe aussi par le développement d'outils d'explication, de justification et de contestation adaptés aux besoins des algorithmes publics. Ces outils ne peuvent aboutir qu'à l'issue de longues phases de recherche, de développement et d'évaluation, y compris en conditions réalistes.

Une double expérimentation menée auprès de la Commission nationale de l'informatique et des libertés (Cnil)

Comme évoqué plus haut, les productions du champ de recherche en XAI ne sont pas toujours adaptées à l'ensemble des problématiques soulevées par la mise en œuvre concrète de SDA. En plus d'être généralement trop techniques (car répondant souvent à un objectif implicite d'amélioration des systèmes par des experts), les outils développés se focalisent sur les explications alors que les justifications jouent un rôle-clé dans la responsabilisation et la légitimation des SDA. Pour évaluer leur pertinence dans un cas d'usage pratique appliqué à l'administration française, nous avons développé deux outils d'explication et de justification qui ont chacun fait l'objet d'une expérimentation auprès d'agents du service chargé des contrôles de la Cnil. Cette étude, basée sur une mise en situation d'audit d'algorithmes, a produit plusieurs résultats.

- Des retours qualitatifs ont été recueillis lors d'une série d'entretiens et grâce à des commentaires libres laissés sur l'interface d'expérimentation. Ces retours ont permis d'identifier les besoins spécifiques en outils d'explications et de justifications pour les audits d'algorithmes et de mieux cerner les modalités d'utilisation de ces outils par des utilisateurs réels.
- Des grandeurs relatives à la performance ont pu être directement mesurées lors de la réalisation de tâches fictives avec (ou sans) les outils afin d'évaluer leur efficacité sur des tâches prédéfinies.

Contexte et déroulement de l'étude

Depuis sa création en 1978, la Cnil a pour mission de veiller à la protection des données personnelles contenues dans les fichiers et traitements, aussi bien publics que privés. Dans ce cadre, elle mène diverses activités, dont le contrôle des entités publiques ou privées réalisant de tels traitements. Lorsqu'un contrôle relève des manquements, la Cnil peut prononcer des mesures ou sanctions à l'égard du responsable de traitement. Pour accomplir cette mission, elle s'appuie sur un savoir-faire qui repose sur une étroite collaboration entre

8 Citation traduite par l'auteur.

auditeurs des systèmes d'information et juristes. Ces contrôles peuvent être réalisés en ligne, sur place ou à partir de documents et peuvent concerner des éléments facilement vérifiables (par exemple, l'oubli d'une mention légale dans les conditions générales d'utilisation) ou des aspects complexes (par exemple, détecter que des données ont été utilisées pour l'entraînement d'un système alors que le responsable n'est pas supposé y avoir accès).

Lorsque le contrôle porte sur un SDA pouvant employer des techniques sophistiquées, les agents de la Cnil s'en tiennent le plus souvent à des entretiens avec les équipes techniques ou commerciales, et, dans certains cas, à une analyse rétrospective des décisions prises par le système. Les agents de la Cnil n'ont pas toujours recours à une analyse de code ni à une analyse en boîte noire pour mener leur contrôle. Cette limitation est principalement liée à des considérations pratiques, les durées d'instruction étant faibles (généralement inférieures à trois jours) alors que les systèmes d'information concernés peuvent être très complexes. Cependant, les agents interrogés précisent que la constatation de l'infraction est souvent simple, et que les entretiens permettent en général d'obtenir une reconnaissance des manquements par les responsables du traitement. Bien que les contrôles actuels ne nécessitent pas systématiquement d'outils spécifiques à l'analyse des algorithmes, l'usage croissant de l'IA dans de nombreux services commerciaux et administratifs ainsi que l'évolution de la réglementation incitent la Cnil à s'intéresser à de tels outils pour conduire ses futurs contrôles et donc à participer à des études expérimentales sur des outils de l'XAI.

En raison de la crise sanitaire liée à l'épidémie du Covid-19, l'étude a été entièrement réalisée en ligne sur une plateforme d'expérimentation[9] et en visioconférence. Avant d'accéder à la plateforme, les participants ont assisté à une brève présentation de l'étude. Afin de ne pas influencer l'utilisation des outils, peu de détails ont été fournis sur la manière de les prendre en main et d'interpréter les informations fournies. En plus de se rapprocher des conditions réelles de déploiement, le caractère sommaire de cette présentation permet d'évaluer également les éventuelles difficultés rencontrées pour se servir concrètement des outils.

L'expérimentation s'est déroulée sous la forme d'une mise en situation où les participants ont endossé le rôle d'agents chargés de réaliser un audit algorithmique sur une entreprise responsable d'un SDA déterminant l'attribution de crédits à la consommation. L'algorithme, fictif et développé pour les besoins de l'expérimentation, se base sur 13 variables comprenant des informations relatives au crédit demandé (pourcentage d'apport, montant, etc.), à l'historique des crédits (autres crédits en cours, défauts, etc.) et à l'âge du demandeur. Ces variables sont issues d'un jeu de données publiques relatives à des crédits à la consommation contenant un grand nombre de décisions prises par les agents de l'établissement de crédit. Pour avoir un algorithme complexe, et donc opaque du point de vue des participants qui cherchent à l'auditer, une IA a été utilisée pour reproduire ces décisions. Le nombre de variables et la complexité du fonctionnement logique du SDA rendent le fonctionnement du système difficile à prévoir, en particulier lorsqu'il est manipulé dans le temps réduit de l'expérimentation. Afin de simuler des politiques d'établissements lisibles, que les participants devront tenter de découvrir, trois règles spécifiques ont été ajoutées au SDA, sans que celles-ci soient rendues visibles aux participants.

- Règle 1 : tous les dossiers dont le demandeur est âgé de plus de 60 ans sont systématiquement refusés (règle volontairement discriminatoire) ;
- Règle 2 : les demandeurs ayant enregistré plus de deux crédits en défaut au cours des six derniers mois sont systématiquement rejetés ;
- Règle 3 : les premiers crédits (aucun crédit en cours) pour des montants faibles (inférieur à un seuil déterminé) sont systématiquement acceptés.

Présentation de l'expérimentation IBEX

L'outil IBEX (*Interactive Black-box EXplanations*) est un outil d'explication en boîte noire qui fournit des informations permettant de comprendre le fonctionnement de l'algorithme (Henin et Le Métayer, 2021a). Les outils d'explication en boîte noire partent du principe que le code de l'algorithme n'est pas connu. Pour inférer des informations sur son fonctionnement, ceux-ci analysent les liens entre les entrées et les sorties du système. Bien que cette démarche en boîte noire limite les explications possibles, elle présente plusieurs avantages pratiques qui justifient son adoption. En premier lieu, l'analyse en boîte noire

9 Le code de la plateforme est publiquement accessible sur le Gitlab d'Inria : projet Algaudit (https://gitlab.inria.fr/chenin/algaudit).

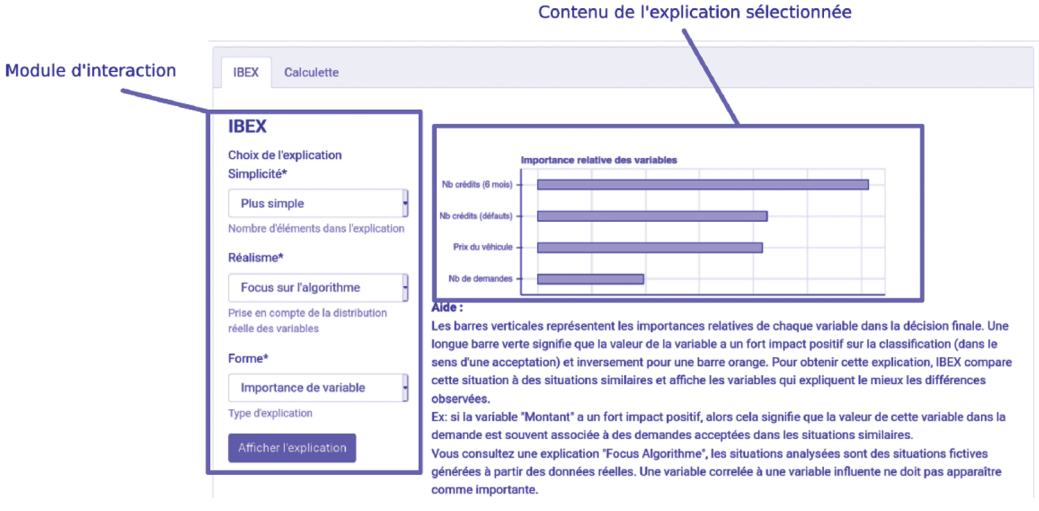
s'applique à de nombreuses situations réelles où le code n'est pas accessible (en raison de la protection de la propriété intellectuelle ou de la complexité des programmes). De plus, cette démarche est indépendante du type d'algorithme considéré, garantissant ainsi une large application et un gain d'efficacité puisqu'une méthode unique peut traiter un grand nombre de systèmes. La particularité d'IBEX réside dans la possibilité offerte à l'utilisateur de choisir la forme et certaines caractéristiques de l'explication qu'il consulte. Le but de la flexibilité offerte aux utilisateurs est de mettre à disposition une diversité d'explications afin de satisfaire aux besoins particuliers de différents utilisateurs (profanes, professionnels, auditeurs, etc.) en fonction d'objectifs différents (améliorer le SDA, contester une décision, vérifier la régularité du système, etc.). Pour cette expérimentation, les participants ont eu accès à une version simplifiée de l'outil d'explication, leur permettant de choisir l'explication en fonction de sa simplicité, de son réalisme et de sa forme (voir Figure 1).

Dans le cadre de l'expérimentation, la plateforme présente une série de décisions du SDA à chaque participant et lui demande de répondre aux questions suivantes :

- quel a été le facteur déterminant de la décision (entrée à choix unique) ?
- d'autres facteurs ont-ils influencé la décision (entrée à choix multiples) ?
- comment comprenez-vous la décision (texte libre) ?

Les deux premières questions visent à évaluer le niveau général de compréhension des participants et la dernière, dont la réponse est sous forme de texte libre, cherche à évaluer la capacité des participants à identifier les règles spécifiques simples (cf. *supra*), et notamment la première, qui est volontairement discriminante.

Figure 1 : capture d'écran de l'outil IBEX sur la plateforme d'expérimentation



À partir des réponses, ont pu être extraites cinq mesures quantitatives pour évaluer l'utilité d'IBEX. Pour disposer d'un point de comparaison, l'ensemble des participants n'a pas eu accès à IBEX : une partie d'entre eux (le groupe témoin) a dû se contenter d'une calculatrice permettant de simuler le SDA. Les performances du groupe IBEX sont comparées à celle du groupe témoin sur les cinq critères suivants :

1. capacité à identifier le « facteur principal » de la décision ;
2. capacité à identifier les autres facteurs de la décision ;
3. capacité d'expliquer la décision sous forme de texte libre ;
4. capacité à distinguer les cas simples, pour lesquels l'algorithme s'appuie sur un grand nombre d'exemples du jeu d'entraînement allant dans le même sens, et les cas limites correspondant à des situations plus équilibrées ;
5. durée nécessaire pour réaliser une tâche.

Résultats et conclusions de l'expérimentation IBEX

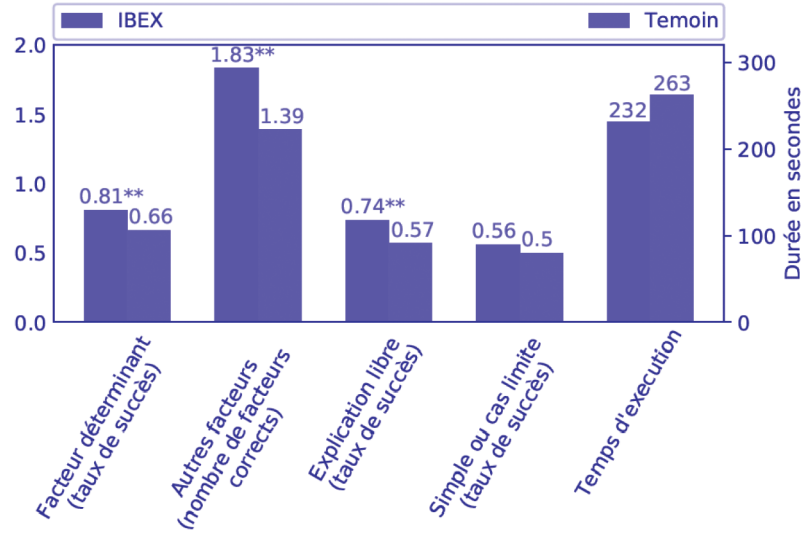
Tout d'abord, les participants, auditeurs du service des contrôles de la Cnil, ont apprécié l'approche boîte noire en comparaison d'une approche boîte

blanche[10] pour la pratique actuelle des contrôles. En effet, bien qu'il soit théoriquement possible d'obtenir le code du système dans le cadre d'un audit, les analyses en boîte noire s'avèrent plus efficaces, car un même outil peut être utilisé sur de nombreux systèmes. De plus, il est difficile de connaître les modalités exactes d'exécution d'un code une fois placé dans un environnement de production. Les explications en boîte blanche peuvent néanmoins s'avérer utiles – et même supérieures aux explications en boîte noire– dans d'autres contextes, notamment pour superviser des algorithmes appliqués à des décisions administratives à forts enjeux ou à grande échelle (Merigoux, Alauzen, Banuls *et al.*, 2024).

Les résultats qualitatifs (obtenus à partir d'entretiens avec questions ouvertes) ont révélé que les participants à l'étude ont jugé le contenu des explications globalement compréhensible et utile. Tous les répondants ayant eu accès à IBEX ont déclaré que l'outil pourrait s'avérer bénéfique pour un audit d'algorithme. La possibilité d'interagir avec l'explication a été perçue comme simple et intuitive, bien que la manière d'utiliser le réalisme de l'explication pour répondre aux questions ait soulevé quelques interrogations. Les participants ont, par exemple, apprécié la possibilité de modifier la forme de l'explication pour avoir différents « angles d'approche » sur l'algorithme et pour confirmer les informations acquises avec les autres formes. Quelques suggestions ont été émises, comme l'intégration d'explications globales (généralement possible avec IBEX, mais non disponible lors de cette expérimentation) ou la mise à disposition d'informations spécifiquement conçues pour les contrôles.

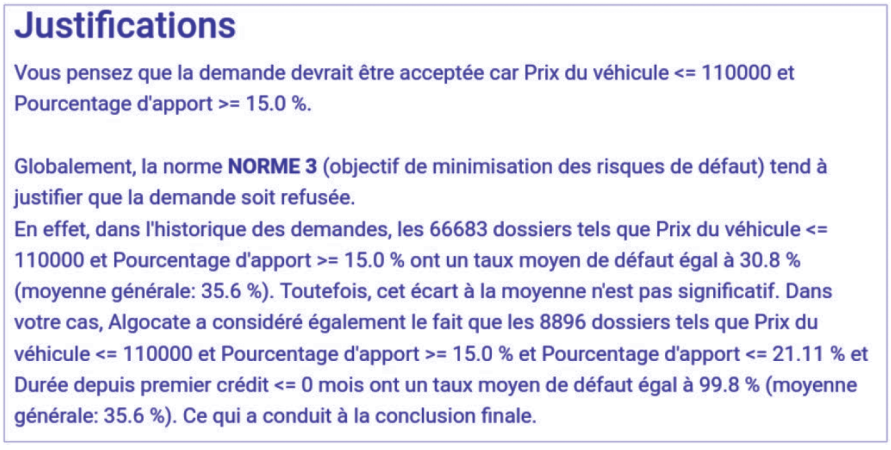
Quantitativement, parmi les cinq critères sélectionnés, IBEX démontre une performance supérieure à l'absence d'outil ; cette différence est significative pour trois critères (voir Figure 2). En particulier, IBEX permet d'identifier le facteur principal dans 81 % des exercices, contre 66 % pour le groupe témoin, et permet d'identifier en moyenne 1,83 autre facteur, par rapport à 1,39 pour le groupe témoin. Ces résultats sont prometteurs et valident l'utilité pratique d'IBEX ainsi que, plus généralement, la démarche en boîte noire. Nous avons également pu comparer la détection de la règle discriminante entre le groupe test et le groupe témoin. Ainsi, un seul participant sur 14 dans le groupe IBEX n'a pas signalé le caractère discriminant de l'algorithme, contre trois sur 15 dans le groupe témoin. IBEX semble donc offrir un avantage, bien qu'un critère plus spécifique (par exemple, une règle complexe combinant plusieurs variables) aurait sans doute été encore plus révélateur.

Figure 2 : IBEX est plus performant sur les cinq critères retenus et la différence est significative (p-valeur < 0.05) sur les trois critères identifiés par « ** ».



10 À l'inverse d'une démarche en boîte noire, cela suppose un accès au code source du système ainsi qu'aux éventuels coefficients du modèle d'IA employé.

Figure 3 : exemple de justification fournie par Algocate sur la plateforme d'expérimentation.



Expérimentation de l'outil de justification Algocate

L'outil Algocate (voir exemple Figure 3) est un dispositif de contestation et de justification des décisions algorithmiques (Henin et Le Métayer, 2021b). Il repose sur un dialogue entre l'utilisateur et l'outil, basé sur des normes justifiant l'utilisation du SDA : respecter une règle simple, poursuivre un objectif quantitatif ou reproduire des décisions jugées valables.

Dans la mise en situation présentée aux participants, l'établissement de crédit déclare que le SDA s'appuie sur trois normes, classées par ordre de priorité, pour prendre les décisions.

1. Norme 1 (règle métier) : pour minimiser les risques de défaut et de surendettement, les personnes ayant connu au moins deux défauts sont systématiquement refusées ;
2. Norme 2 (règle métier) : pour faciliter l'accès au marché aux nouveaux clients, les personnes sollicitant un crédit pour la première fois sont acceptées, quelles que soient les caractéristiques de leur demande, à condition que le prix du bien soit inférieur à un seuil déterminé ;
3. Norme 3 (objectif) : les risques de défaut de crédit doivent être réduits. Une base de données historiques de demandes de crédit est utilisée pour estimer les risques associés à une nouvelle demande.

Les deux premières normes correspondent aux règles métiers déjà présentées, et la dernière indique que les décisions s'appuient sur une base de données pour minimiser les risques. Les règles du jeu sont donc inversées par rapport à l'expérimentation précédente. Dans le premier cas, on cherche à déterminer le fonctionnement alors que, dans le second, ce fonctionnement est connu et on cherche à vérifier qu'il est appliqué dans des situations spécifiques. La règle discriminante n'est pas déclarée par l'établissement. À nouveau, la plateforme présente une série de décisions et le participant est invité à utiliser l'outil Algocate et à répondre aux questions suivantes :

- La décision vous semble-t-elle justifiée au regard des normes déclarées par la société (oui ou non) ?
- Indiquez le niveau de confiance de votre réponse (sur une échelle de 1 à 5).
- Justifiez votre choix (sous forme de texte libre).

Afin d'introduire artificiellement des décisions « injustifiées », une décision sur deux est délibérément modifiée de façon à ce que cette dernière ne reflète plus les normes de l'entreprise.

Des résultats mitigés pour l'outil Algocate qui ne remettent pas en cause l'intérêt des justifications

La grande majorité des utilisateurs a rapidement intégré les notions de normes, de contestation et de justification. Une partie des répondants estime que l'utilisation de normes dans le cadre de contrôles ou d'audits algorithmiques est éclairante, car elle permet de confronter directement le responsable à ses déclarations. Toutefois, les participants ont globalement trouvé l'outil complexe à utiliser, notamment par rapport à IBEX, et ont souligné qu'il n'était actuellement pas adapté aux contrôles dans la mesure où les responsables de SDA n'explicitent pas leurs normes comme l'exigerait Algocate. Concernant l'utilisation de l'outil, si certains répondants ont rencontré des difficultés à

comprendre l'interface, la forme des justifications, et particulièrement l'usage des statistiques qui rendent l'argument à la fois clair et neutre, ont été unanimement appréciés. En ce qui concerne les normes, elles ont été globalement bien comprises bien que celle sous forme d'objectif ait suscité plus d'interrogations. Celle-ci a été qualifiée de norme « floue », car elle peut être utilisée pour soutenir ou contester une même décision selon l'interaction avec l'utilisateur. Selon nous, ces réflexions semblent pertinentes, car elles reflètent une réalité où certains sous-ensembles du jeu de données soutiennent un refus tandis que d'autres soutiennent une acceptation de la décision.

En ce qui concerne les performances des utilisateurs, ils ont été en mesure de détecter les décisions injustifiées dans 65 % des tâches. Étant donné que certaines tâches ont été réalisées sans recourir à Algocate, nous avons pu décomposer les résultats entre deux groupes. Il apparaît que les participants utilisant cet outil ont obtenu de meilleurs résultats (67 % contre 61 %), bien que la différence ne soit pas significative. La confiance autodéclarée (sur une échelle de Likert de 1 à 5) indique qu'Algocate améliore globalement la confiance des participants, surtout lorsque ces derniers ont fourni la bonne réponse.

Pour résumer, cette expérimentation apporte des enseignements sur l'utilisation du protocole Algocate. Tout d'abord, elle valide la possibilité de l'implémenter et de le mettre à disposition d'utilisateurs non spécialistes. Même si l'interface a été jugée complexe, les principes fondamentaux ont généralement été compris rapidement par les participants. Ces résultats préliminaires ne permettent toutefois pas de démontrer de façon concluante qu'Algocate est efficace en conditions réelles pour détecter des décisions ne répondant pas à un ensemble de normes prédéfinies.

Quels que soient les résultats précis de ces expérimentations, l'intérêt général des méthodes d'explication et de justification semble démontré. L'expérimentation, menée avec des professionnels de l'audit externe, met en lumière la nécessité pour les administrations de se mettre en capacité de maîtriser totalement les SDA qu'elles déploient et d'en démontrer l'adéquation avec les objectifs assignés à ces systèmes (à supposer que ces derniers aient été clairement définis). Au-delà des innovations techniques qu'ils représentent, ces outils d'explication et de justification visent à ramener les SDA dans un espace de délibération sur leurs finalités et leurs implications en participant à effacer les aspects techniques, complexes et opaques. Les parties concernées par ces systèmes ont des niveaux d'implication et d'expertise très divers allant du directeur d'administration, responsable des décisions opérées par le système, aux personnes affectées par ces dernières en passant par les agents en charge de la mise en œuvre opérationnelle. Dans un contexte de passage à l'échelle des SDA, il faut disposer des outils permettant de faire la traduction entre l'ensemble de ces parties prenantes et les enjeux techniques pour légitimer et donc gagner la confiance des acteurs et citoyens impliqués.

Conclusion

La digitalisation de l'action publique et le développement des méthodes d'analyse des données constituent une opportunité pour l'administration de développer et de mettre en œuvre un nombre croissant d'algorithmes pour assister ou prendre certaines décisions. Ces algorithmes pourraient améliorer les processus de décision en les rendant plus efficaces, transparents et objectifs à condition que les systèmes eux-mêmes le soient. Pour que le développement des SDA soit salutaire, il nous paraît essentiel que les débats autour de leur conception portent à la fois sur les aspects techniques, indispensables à leur mise en œuvre, et sur les objectifs qui leur sont assignés. Ces derniers, que nous appelons dans cet article des « normes » et sur lesquels repose la légitimité des décisions, devraient systématiquement faire l'objet d'une formalisation et d'une communication claires auprès des agents impliqués, des personnes concernées par les décisions et auprès des régulateurs. Alors que l'essor des SDA au sein de l'administration pourrait s'accélérer dans les années et les décennies à venir, ces considérations sont déjà actuelles. À titre d'exemple, en France, la Cour des comptes reprochait en 2024 à la direction générale des Finances publiques un défaut de communication sur les logiques utilisées par leurs algorithmes de ciblage pour les contrôles, et recommandait de se doter d'une stratégie plus claire[11]. Afin de favoriser un dialogue plus équilibré entre les SDA et les personnes affectées par les décisions, de faciliter

11 « La formalisation d'une stratégie de détection des irrégularités fiscales permettrait d'expliciter la place réservée à chacune de ces logiques et de répartir les moyens affectés au contrôle en fonction d'objectifs transparents, clairement formulés et affichés. » (Cour des comptes, 2024)

la supervision et la régulation de ces systèmes et d'assurer que leur développement se fait bien au bénéfice des administrations et des usagers, des outils fournissent des informations pour expliquer ou justifier le fonctionnement des SDA. Ces outils, qui paraissent indispensables au développement des SDA, doivent rester proches des usages réels et faire l'objet d'évaluations rigoureuses avec des utilisateurs humains et, autant que possible, dans des contextes réalistes.

Clément Henin est docteur en informatique de l'Inria et a réalisé sa thèse sur l'explication et la justification des systèmes algorithmiques de décision. Il est actuellement directeur du service de science des données à l'Assistance publique – Hôpitaux de Paris.

Références bibliographiques

Binns R. (2018),
"Algorithmic Accountability and Public Reason", *Philosophy & Technology*, vol. 31, n° 4, pp. 543-556.

Commission européenne, Direction générale des réseaux de communication, du contenu et des technologies (2019),
Lignes directrices en matière d'éthique pour une IA digne de confiance, Publications Office, https://data.europa.eu/doi/10.2759/74304.

Cour des comptes (2024),
L'action de la direction générale des finances publiques auprès du bloc communal, janvier, https://www.ccomptes.fr/fr/publications/laction-de-la-direction-generale-des-finances-publiques-aupres-du-bloc-communal

Henin C. (2021),
« Confier une décision vitale à une machine », *Réseaux*, n° 225(1), pp. 187-213.

Henin C., Le Métayer D. (2021a),
A Multi-layered Approach for Tailored Black-Box Explanations. ICPR International Workshops and Challenges 2021. *Lecture Notes in Computer Science*, Springer vol. 12663.

Henin C., Le Métayer D. (2021b),
A framework to contest and justify algorithmic decisions. AI&Ethics 1, pp. 463-476.

Hildebrandt M. (2019),
"Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning", *Theoretical Inquiries in Law*, vol. 20, pp. 83-122 (spéc. p. 118).

Kluttz D.N., Kohli N. et Mulligan D.K. (2020),
"Shaping our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions", *in* Werbach K. (ed.), *After the Digital Tornado. Networks, Algorithms, Humanity*, Cambridge University Press, pp. 137-152.

Merigoux D., Alauzen M., Banuls J *et al.* (2024),
De la transparence à l'explicabilité automatisée des algorithmes : comprendre les obstacles informatiques, juridiques et organisationnels. RR-9535, INRIA Paris, 68 pages, https://inria.hal.science/hal-04391612v1.

Miller T. *et al.* (2017),
"Explainable AI : Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences", ArXiv abs/1712.00547.

Porter Theodore M. (1996),
Trust in numbers: The pursuit of objectivity in science and public life, Princeton University Press.

Rouvroy A. (2013),
"The End(s) of Critique: Data Behaviourism versus Due Process", *Privacy, Due Process and the Computational Turn*, Routledge, pp. 157-182.

Suchman M.C. (1995),
"Managing Legitimacy: Strategic and Institutional Approaches", *Academy of Management Review*, vol. 20, pp. 571-610.

La régulation de l'intelligence artificielle aux États-Unis

Winston Maxwell

Aux États-Unis, l'usage de l'IA passe par plusieurs strates de régulation, posant régulièrement des questions d'harmonisation législative selon les différents territoires. Dans ce contexte, et alors même que les États fédérés sont et resteront des acteurs importants de cette régulation, l'arrivée d'une nouvelle administration susceptible d'orienter la politique vers une certaine dérégulation présente un risque de déséquilibre entre les niveaux de législation. Si la question de la prééminence technologique des États-Unis reste une priorité nationale, la généralisation de l'IA au niveau administratif posera certainement question. Entre usage efficient et limites indispensables, cet article discute du véritable intérêt de cette technologie à l'échelle d'une démocratie si complexe.

La régulation de l'IA aux États-Unis agit en plusieurs strates : la régulation au niveau fédéral, la régulation au niveau des États fédérés, et la régulation au niveau des municipalités. La structure fédérale des États-Unis accorde une grande liberté de légiférer aux États fédérés, ce qui a conduit naturellement à la multiplication de lois sur l'IA : biométrie, IA dans la gestion des ressources humaines, protection de la vie privée, hyper-trucage, profilage publicitaire, risques liés aux très grands modèles de langage (*Large Language Models*, LLM). Cette activité de régulation est même descendue au niveau des municipalités, qui ont par exemple utilisé leurs pouvoirs locaux pour réguler l'usage de l'IA dans le domaine de l'emploi[1]. Au niveau fédéral, l'action de régulation est plus limitée, en partie à cause de la situation politique divisée au Congrès. Celui-ci s'est focalisé sur la protection du *leadership* technologique des États-Unis en IA et sur les risques de l'IA pour la sécurité nationale, des sujets qui rassemblent sur le plan politique. Certaines lois fédérales anciennes, notamment celles sur la protection du consommateur et les lois contre les discriminations, sont mobilisées au niveau fédéral pour combattre les usages abusifs de l'IA, mais sans disposition spécifique visant les risques d'IA.

La coexistence de différentes couches de lois sur l'IA pose un problème d'harmonisation au niveau fédéral. Une application de biométrie devra, par exemple, obéir à des normes différentes selon qu'elle est déployée dans l'État d'Illinois, ou dans la ville de San Francisco. Ce patchwork de normes peut surprendre, mais il est fréquent aux États-Unis. Chaque État fédéré dispose, par exemple, de sa propre loi sur la protection des données à caractère personnel. Dans ce contexte de décentralisation, certains États, tels que la Californie, jouent le rôle de *leader*. On parle même d'« effet Californie » dans le domaine des lois sur l'environnement, car les normes fixées dans cet État peuvent se répandre à d'autres[2]. Cette diversité normative peut créer des tensions, mais permet également d'expérimenter différentes approches de régulation (Roesler, 2014 ; Whitt, 2009 ; Livermore, 2016).

L'élection de Donald Trump, soutenue par une majorité républicaine au Congrès, ouvrira un nouveau chapitre de régulation, ou plutôt de dérégulation, de l'IA au niveau fédéral. Le vice-président J.D. Vance, et le futur directeur du ministère de l'efficacité gouvernementale Elon Musk, ainsi que leurs soutiens dans la Silicon Valley, épousent le principe de « techno-solutionnisme », à savoir que la technologie peut, mieux que la régulation, résoudre la plupart des problèmes de

1 https://rules.cityofnewyork.us/wp-content/uploads/2023/04/DCWP-NOA-for-Use-of-Automated-Employment-Decisionmaking-Tools-2.pdf

2 https://www.iatp.org/sites/default/files/Environmental_Regulation_and_Economic_Integrat.pdf

l'humanité. Le techno-solutionnisme peut, selon certains, devenir une menace pour la démocratie (Nemitz, 2023 ; Lafrance, 2024).

La nouvelle administration Trump pourrait agir sur trois axes en matière de régulation de l'IA. D'abord, elle chercherait à renforcer les lois visant à préserver la prééminence technologique des États-Unis par rapport à la Chine. Ensuite, l'administration et le Congrès pourraient tenter de détricoter une partie de la régulation de l'IA, particulièrement celle des très grands modèles dits « de frontière », au niveau des États fédérés. Enfin, la nouvelle administration tentera d'augmenter l'utilisation de l'IA par les services de l'État pour les rendre plus efficaces. En charge de l'efficacité du gouvernement, Elon Musk souhaitera probablement introduire l'IA dans tous ses échelons.

Cet article examinera ces trois tendances en se concentrant d'abord sur les aspects de la régulation américaine qui visent à assurer la suprématie technologique des États-Unis en matière d'IA, un cadre déjà amorcé par la première administration Trump en 2019, et poursuivi par l'administration Biden. Ensuite, l'étude portera sur les lois et règlements américains qui visent à protéger les citoyens contre les risques de l'IA, et en particulier les nombreux textes adoptés par les États fédérés et les municipalités pour combler le vide laissé au niveau fédéral. Enfin, sera évoquée l'utilisation de l'IA par les administrations et les tribunaux américains, une tendance qui risque de créer des tensions avec la Constitution américaine, que la nouvelle administration Trump est susceptible d'accentuer.

Une réglementation pour protéger la prééminence technologique des États-Unis en matière d'IA

La première administration Trump a adopté en 2019 une directive administrative (*executive order*)[3] pour promouvoir le *leadership* des États-Unis en matière d'IA. Cette directive a été suivie par une loi, le « National Artificial Administration Initiative Act » de 2020[4], visant à coordonner l'ensemble de la politique en matière d'IA au sein des administrations fédérales et à encourager la recherche. Les actions entreprises ensuite par l'administration Biden s'appuient en partie sur cette loi de 2020, adoptée pendant les derniers jours du premier mandat Trump. Le congrès a adopté ensuite le « CHIPS and Science Act » de 2022[5] prévoyant des investissements massifs dans la recherche de pointe, notamment dans la computation quantique, et dans le développement de sites industriels et de compétences pour fabriquer les semiconducteurs aux États-Unis, afin de réduire la dépendance des États-Unis à l'égard de Taiwan. Certains processeurs GPU du fabricant NVIDIA sont dorénavant soumis à des contrôles d'exportation[6]. L'administration Biden a adopté en 2023 une directive administrative (*executive order*)[7] visant à renforcer encore le *leadership* mondial des États-Unis en matière d'IA, tout en protégeant les citoyens contre ses dérives. Cette directive a été complétée le 24 octobre 2024 par une circulaire qui vise spécialement la sécurité nationale[8]. Ces mesures permettent notamment d'assouplir les conditions de visa pour attirer aux États-Unis les meilleurs talents mondiaux en matière d'IA. La directive Biden prévoit l'obligation de déclarer à l'administration les modèles les plus puissants qui peuvent être à double usage, civil et militaire, pour lui permettre de connaître leur existence et d'évaluer la nécessité de mettre en place des contrôles à l'exportation. Cette démarche vise aussi à garantir que les forces armées américaines auront accès aux modèles les plus puissants. Les entreprises américaines fournissant de grandes infrastructures de calculs devront déclarer l'existence de clients étrangers cherchant à entraîner de très grands modèles sur ces infrastructures. La directive Biden de 2023 prévoit la mise en place des ressources computationnelles (*compute clusters*) et des données d'apprentissage massives pour favoriser la recherche, y compris par les PME, ainsi que des programmes de formation en IA au sein de l'administration. Ces mesures s'inscrivent dans la continuité des lois de 2020 (« National Artificial Intelligence Research Act ») et 2022 (« CHIPS Act ») pour garantir la prééminence

3 https://trumpwhitehouse.archives.gov/articles/accelerating-americas-leadership-in-artificial-intelligence/
4 https://www.congress.gov/bill/116th-congress/house-bill/6216
5 https://www.congress.gov/bill/117th-congress/house-bill/4346
6 https://www.reuters.com/technology/nvidia-may-be-forced-shift-out-some-countries-after-new-us-export-curbs-2023-10-17/
7 https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/
8 https://www.whitehouse.gov/briefing-room/presidential-actions/2024/10/24/memorandum-on-advancing-the-united-states-leadership-in-artificial-intelligence-harnessing-artificial-intelligence-to-fulfill-national-security-objectives-and-fostering-the-safety-security/

des acteurs américains dans le développement de l'intelligence artificielle. Même si le candidat Trump a annoncé pendant sa campagne qu'il abrogerait les mesures adoptées par l'administration Biden, il est probable que les mesures visant à renforcer la souveraineté numérique des États-Unis, et notamment l'interdiction d'exporter certaines technologies vers la Chine, restent en vigueur.

L'administration Biden inclut dans sa conception du *leadership* américain un rôle dans les discussions multilatérales sur l'IA au sein de l'ONU, de l'OCDE, et des organisations de normalisation. Cette approche vise à exporter vers le monde non américain les normes techniques, et les approches de gouvernance et de régulation d'IA développées aux États-Unis, notamment au sein de l'agence « National Institute of Standards and Technology » (NIST). Celle-ci développe des protocoles de test[9] et des normes de gouvernance[10] de l'IA qui font dorénavant référence au niveau mondial. L'agence NIST héberge l'IA Safety Institute[11], créé pour développer des normes de sécurité pour l'IA et faciliter des échanges à l'international. Il s'agit d'utiliser le *soft power* américain pour encourager une approche mondiale compatible avec l'approche américaine de la régulation de l'IA, et donc favorable aux acteurs américains. Compte tenu de l'hostilité du futur président Trump aux discussions multilatérales, il est probable que ce volet de la stratégie IA de l'administration Biden soit dépriorisé, même si l'administration Trump partage l'objectif de rendre le cadre de régulation mondial compatible avec les intérêts des grands acteurs américains de l'IA.

Les mesures de régulation visant à protéger les citoyens contre les risques de l'IA

Au-delà du sujet du *leadership* technologique, la régulation de l'IA doit naturellement se pencher sur les risques pour les citoyens. Ceux-ci concernent les discriminations algorithmiques, l'incapacité des humains à bien comprendre et maîtriser ces systèmes, les menaces pour l'emploi, voire des risques existentiels soulevés par certains scientifiques, dont le prix Nobel Geoffrey Hinton (Heaven, 2023).

Faute de loi spécifique sur les risques de l'IA, l'administration fédérale applique les lois générales sur la protection des consommateurs

La directive Biden de 2023 inclut des mesures visant à protéger les citoyens contre les dérives algorithmiques : discriminations, utilisation de l'IA par les tribunaux et la police, menaces pour la vie privée, impacts sur l'emploi, automatisation excessive des fonctions de l'administration. En l'absence d'une loi fédérale qui vise spécifiquement la protection des individus contre les risques de l'IA, la directive Biden n'a de valeur contraignante qu'à l'égard de l'administration fédérale. À l'instar de l'« AI Act » européen, la directive Biden impose aux différentes administrations l'obligation de préparer des analyses des risques, notamment à l'égard les populations vulnérables, de mettre en place des programmes de gestion des risques, et notamment des audits, et de créer des instances de gouvernance, dont un « Chief AI Officer » au sein de chaque administration. Les acteurs privés sont indirectement impactés par la directive Biden dans la mesure où ils fournissent des systèmes à l'administration fédérale. De plus, la directive Biden demande aux administrations d'appliquer les lois existantes, de portée générale, aux utilisations de l'IA, par exemple le « Federal Trade Commission Act » (« FTC Act ») qui interdit des pratiques déloyales ou trompeuses dans le commerce. Cette loi permet à la FTC de poursuivre des entreprises qui utilisent l'IA de manière déloyale contre les citoyens[12]. La FTC utilise également le droit de la concurrence, comme en témoigne l'enquête lancée contre OpenAI, Microsoft et NVIDIA[13].

Le « Fair Credit Reporting Act » de 1970 (FCRA)[14] s'applique déjà à toute activité consistant à générer un score ou rapport de solvabilité (*credit score*). Cette loi s'adapte parfaitement à l'IA. Elle impose des obligations de transparence aux entreprises qui préparent ou utilisent des rapports et scores pour prendre des décisions d'octroi de crédits, de location de biens immobiliers, ou de

9 Notamment les tests de biais pour la reconnaissance faciale : https://www.nist.gov/programs-projects/face-technology-evaluations-frtefate
10 Notamment le AI Risk Management Framework : https://csrc.nist.gov/projects/risk-management/about-rmf
11 https://www.nist.gov/aisi
12 La FTC vient déposer une plainte contre une entreprise qui utilise l'IA pour détecter des armes à feu en raison des affirmations trompeuses de l'entreprise sur la fiabilité du système (https://www.ftc.gov/system/files/ftc_gov/pdf/EVOLVCOMPLAINTFILED.pdf), ainsi qu'une plainte contre une entreprise qui propose les services d'un « robo-avocat » par IA.
13 https://www.usine-digitale.fr/article/antitrust-microsoft-nvidia-et-openai-dans-le-viseur-des-autorites-americaines.N2214429
14 https://en.wikipedia.org/wiki/Fair_Credit_Reporting_Act

recrutement ou de licenciement de salariés. La personne visée par le score dispose d'un droit d'accès à celui-ci et un droit de rectification. La loi impose l'obligation de fournir des explications pour une décision prise sur le fondement d'un tel score. Cette obligation est similaire à celle prévue par l'article 86 de l'« AI Act » européen.

L'« Equal Credit Opportunity Act » de 1974 (ECOA)[15] ainsi que le « Civil Rights Act » de 1964[16] interdisent toute forme de discrimination. La FTC est en charge de l'application de l'ECOA et a précisé qu'en application de cette loi, une discrimination indirecte, par exemple un score différencié fondé sur le code postal de l'individu, pourrait tomber sous le coup de la loi[17].

L'application de ces lois à portée générale dépend de la politique d'enquête et de sanction de la FTC, actuellement présidée par Lina Khan, une farouche critique de la Big Tech. Le changement d'administration pourrait conduire à la nomination d'un nouveau président ou présidente de la FTC, plus souple à l'égard des géants de la Silicon Valley.

Ainsi, les États-Unis ne disposent pas d'une loi fédérale spécifique visant à protéger les citoyens contre les dérives de l'IA. Pour trouver des lois qui visent spécifiquement l'IA, il faut descendre au niveau des États fédérés et des municipalités. À majorité démocrate, l'État de Californie a été le plus actif dans l'adoption de lois sur l'IA, même si de nombreux autres États et villes ont également légiféré. La Californie occupe une place particulière parce que les acteurs de la Big Tech ne peuvent se passer de ce marché. Une régulation californienne peut ainsi devenir, *de facto*, la norme nationale, voire internationale, dans la mesure où les fournisseurs de systèmes ne vont pas développer et maintenir plusieurs versions de leurs produits pour le marché américain.

En l'absence d'une loi fédérale, les États fédérés adoptent des lois pour protéger les citoyens contre les dérives de l'IA

Cette activité au niveau de l'État de Californie n'est pas toujours au goût de l'administration fédérale. Pendant la première administration Trump, le gouvernement fédéral a essayé de bloquer l'application d'une loi californienne sur la neutralité de l'Internet en prétendant que le sujet relevait de la compétence exclusive de l'État fédéral. Cette tentative a échoué[18], mais il n'est pas exclu que la deuxième administration Trump tente de bloquer l'application de certaines lois californiennes sur l'IA pour les mêmes raisons. Pour ce faire, il faudrait que le congrès adopte une loi fédérale de régulation de l'IA en précisant spécifiquement que la loi fédérale se substitue à toute législation sur l'IA des États fédérés. L'administration dispose d'une majorité dans les deux chambres du Congrès, donc une telle loi fédérale pourrait éventuellement voir le jour, même si cela prend du temps, et ne peut pas couvrir tous les cas d'usage d'IA. En application de la Constitution, certains domaines restent sous la compétence de chaque État fédéré, même si la frontière de compétence entre l'État fédéral et les États fédérés reste discutée.

Les lois récemment adoptées par la Californie convergent sur de nombreux points avec l'« AI Act » européen. Les lois californiennes[19] imposent des mesures de transparence sur l'utilisation de l'IA, notamment l'utilisation de l'IA générative pour modifier ou créer des contenus. Des mesures de transparence incluent l'obligation de publier des informations sur les données d'apprentissage, y compris la présence de données à caractère personnel ou de contenus protégés par le droit d'auteur. La loi qui protège les données émises par le cerveau vise les neuro-technologies et notamment la société Neuralink d'Elon Musk. La Californie a également légiféré pour protéger l'image des acteurs, artistes et autres personnes publiques. Un projet de loi en discussion en Californie obligerait les fournisseurs de très grands modèles – appelés « modèles de frontière » – à effectuer une analyse des risques systémiques pour la sécurité, et mettre en place des mesures pour atténuer ces risques. Les négociations sur le règlement « AI Act » ont montré que les acteurs de la Big Tech sont hostiles à une régulation de ces très grands modèles. On peut donc imaginer que l'administration Trump, dorénavant alliée avec une partie des acteurs de la Silicon Valley, tentera de freiner les ambitions de la Californie pour les réguler.

L'État de Washington était précurseur dans la régulation de la reconnaissance faciale, en imposant

15 https://en.wikipedia.org/wiki/Equal_Credit_Opportunity_Act
16 https://en.wikipedia.org/wiki/Civil_Rights_Act_of_1964
17 https://www.ftc.gov/business-guidance/blog/2020/04/using-artificial-intelligence-algorithms
18 https://www.theverge.com/2022/1/28/22906856/court-upholds-california-net-neutrality-law
19 https://www.gov.ca.gov/2024/09/29/governor-newsom-announces-new-initiatives-to-advance-safe-and-responsible-ai-protect-californians/

dès 2020 un encadrement strict, et notamment des tests de biais, analyses de risques, mesures de gouvernance et de contrôle humain, pour toute utilisation de la reconnaissance faciale par la police[20]. Les dispositions de cette loi ressemblent à celles de l'« AI Act » (Maxwell, 2023). L'utilisation de la reconnaissance faciale par la police a été interdite dans des dizaines de municipalités telles que San Francisco et Portland, même si la tendance à interdire ces technologies s'est ralentie[21]. Les États de l'Illinois[22] et du Texas[23] ont légiféré sur l'utilisation de systèmes biométriques par des entreprises privées. La ville de New York[24] a adopté un arrêté ambitieux sur l'utilisation de l'IA pour le recrutement et la gestion des ressources humaines dans la ville, même par des entreprises privées.

La constitution américaine impose des limites à l'utilisation de l'IA par les administrations de l'État

Adopté au niveau fédéral, l'« Advancing American AI Act » de 2022[25] encourage de nouveaux programmes pour tester l'utilisation de l'IA au sein des administrations pour rendre le fonctionnement du gouvernement plus efficace. Cette tendance sera sans doute amplifiée sous l'influence du futur ministre de l'efficacité gouvernementale, Elon Musk. L'utilisation de l'IA par les administrations et tribunaux, que ce soit au niveau fédéral ou au niveau de chaque État fédéré, peut néanmoins se heurter à deux dispositions de la Constitution américaine. La première est la garantie d'une procédure régulière (*due process*) ; la deuxième est la protection contre une intrusion excessive du gouvernement dans la vie privée (le 4e amendement).

Le principe constitutionnel de « due process » limite le recours à l'IA par les administrations

Le principe constitutionnel de *due process* garantit qu'une personne ne sera pas privée d'un droit sans une procédure régulière. Garanti par les 5e et 14e amendements de la Constitution, ce principe comporte deux volets : un volet concernant la qualité de la norme sur le fond (*substantive due process*), et un volet de procédure (*procedural due process*). Ce volet procédural permet aux tribunaux de censurer toute décision prise par l'État ou par un tribunal qui ne se conformerait pas aux règles d'une procédure équitable. Le volet procédural est particulièrement mobilisé à l'égard de décisions prises sur le fondement de résultats algorithmiques. En particulier, il garantit que l'utilisation de l'IA par les administrations et tribunaux ne pourra pas avoir pour effet de priver un individu de son droit de comprendre et de contester une décision, et de bénéficier d'une audience (*hearing*) pour faire entendre son point de vue (Maxwell, 2022a). Un algorithme utilisé au Texas pour évaluer les performances des enseignants dans les écoles publiques a été épinglé pour manque de transparence en violation du principe constitutionnel de *due process*. Pour les juges, l'algorithme de performance a été considéré comme l'équivalent d'un test antidopage pour lequel la loi exige la réplicabilité des résultats, notamment pour vérifier l'absence d'erreurs. L'algorithme au Texas ne permettait pas une réplicabilité exacte des résultats, et a donc été jugé anticonstitutionnel[26].

Le constitutionnaliste Aziz Huq a posé la question de savoir si les garanties de *due process* exigeaient systématiquement le recours à un décisionnaire humain, ou si, dans certains cas mineurs, une décision peut être prise par un robot (Huq, 2020a). Il a conclu que rien n'interdisait le recours à un robot pour certaines décisions mineures. Le principe de *due process* garantit la possibilité d'être entendu lors d'une audience (*hearing*). Mais la définition même d'une « audience » reste débattue. Selon Huq et d'autres auteurs (Sunstein, 2022), les juges humains sont biaisés, et un juge automatique pourrait être envisagé en première instance si l'enjeu du litige restait mineur, à l'instar du programme de justice automatique envisagé initialement en Estonie (Park, 2020, pp. 46-48). Pour d'autres auteurs (Brennan-Marquez et

20 https://app.leg.wa.gov/RCW/default.aspx?cite=43.386&full=true
21 https://www.technologyreview.com/2023/07/20/1076539/face-recognition-massachusetts-test-police/
22 https://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=3004
23 https://statutes.capitol.texas.gov/docs/bc/htm/bc.503.htm
24 https://rules.cityofnewyork.us/wp-content/uploads/2023/04/DCWP-NOA-for-Use-of-Automated-Employment-Decisionmaking-Tools-2.pdf
25 https://uscode.house.gov/view.xhtml?req=(title:40% 20section : 11301%20edition:prelim)
26 *Houston Federation of Teachers v. Houston Ind't School Dist.*, 251 F. Supp. 3d 1168 (S.D. Tex. 2017). Voir aussi : Paige, 2020 ; Crawford et Schultz, 2019 ; Maxwell, 2022b.

Henderson, 2019), la présence d'un juge humain reste indispensable en raison des liens essentiels de confiance et d'empathie qui existent lorsqu'un humain est jugé par un autre humain.

Avec le concours d'Elon Musk, la nouvelle administration Trump pourrait être tentée de rendre l'administration plus « efficace », notamment par l'utilisation accrue de l'intelligence artificielle. L'utilisation de l'IA par l'administration américaine a déjà nourri de nombreux débats parmi les spécialistes du droit (Crawford et Schultz, 2019 ; Coglianese et Lehr, 2017 ; Berman, 2018 ; Huq, 2020b). Certains plaident en faveur de l'utilisation de ces outils pour réduire les délais de procédure pour le citoyen, car ceux-ci peuvent eux-mêmes constituer un déni de justice. Les défenseurs de l'IA mettent également en avant le fait que les décisionnaires humains sont souvent plus biaisés que les algorithmes, des tests ayant prouvé que les juges condamnent à des peines plus sévères avant l'heure du déjeuner (Danziger *et al.*, 2011). Les pires injustices sont commises par des humains. L'IA serait en réalité moins discriminatoire que les décideurs humains. Les outils de prédiction statistique concernant le risque de récidive ont été créés justement pour contrer le caractère arbitraire, voire raciste, de décisions humaines sur le traitement de personnes accusées de crimes ou délits (Beaudoin et Maxwell, 2023). Ces outils étaient promus à l'origine par des organismes de défense des droits des prisonniers pour améliorer l'objectivité des processus. Pourtant, une controverse est apparue autour de l'outil COMPAS, déclenchant une polémique sur les discriminations opérées par ces outils, où le taux d'erreur peut varier énormément selon la couleur de peau de l'individu. Le « cas COMPAS » a également provoqué une prise de conscience que tout outil algorithmique sera nécessairement discriminatoire sous au moins un angle, les différents critères d'équité étant incompatibles entre eux (Chouldechova, 2017). Un prisonnier a contesté l'usage de l'algorithme de prédiction sur le fondement du principe de *due process*. La Cour suprême de l'État du Wisconsin a admis qu'un score algorithmique pouvait être utilisé par un juge en tant qu'élément supplémentaire et accessoire d'une décision, mais qu'un tel score ne pourrait jamais constituer un élément déterminant pour la prise d'une décision[27].

Enfin, l'IA peut tout simplement améliorer l'efficacité de l'action publique, par exemple en prédisant les lieux et les heures de criminalité afin d'y affecter davantage d'agents au moment opportun. La « police prédictive » peut cependant nourrir une boucle de rétroaction, car si l'algorithme mobilise plus d'agents dans un quartier classé « à risque », l'augmentation du nombre d'agents fera automatiquement grimper le nombre d'arrestations, ce qui sera interprété par l'algorithme comme un signe de dangerosité accrue du quartier, ce qui déclenchera une présence encore plus forte d'agents, plus d'arrestations, etc.[28] Pendant ce temps, les autres quartiers connaîtront l'inverse : moins d'agents, donc moins d'arrestations, donc moins de violence constatée. L'algorithme amplifie les différences déjà constatées.

L'IA a le potentiel de transformer profondément le fonctionnement de l'État, de nombreuses fonctions de l'État pouvant s'analyser comme celles d'une plateforme (Bertholet et Létourneau, 2017). L'IA peut conduire à une réduction des délais, du nombre d'erreurs, et à plus de justice. Cependant, ceux qui contestent l'utilisation de ces outils soulignent la déshumanisation des relations entre le citoyen et l'administration, et la tendance pour les personnes de l'administration à accorder de plus en plus de confiance à ces outils, rendant le contrôle humain illusoire. Les biais humains de l'automatisation rendent les humains de moins en moins critiques des recommandations algorithmiques. De plus, les humains perdent très vite leurs compétences si elles ne sont plus utilisées tous les jours. Enfin, l'automatisation de l'État amorce un chemin vers un éventuel gouvernement par algorithme, loin du principe du « gouvernement du peuple, par le peuple et pour le peuple » souhaité par le Président Abraham Lincoln en 1863[29]. Un article du professeur Kevin Werbach dresse un parallèle entre la quête d'une meilleure efficacité au sein de l'administration, et le projet chinois de crédit social, décrié en Occident car mettant en place une surveillance quasi orwellienne (Werbach, 2022). Werbach souligne qu'à l'origine ce système chinois cherche à rendre le fonctionnement de son administration plus efficace et plus juste en rendant les citoyens moins dépendants de décisionnaires humains locaux, souvent biaisés et corrompus. La thèse de Werbach est qu'il n'existe pas de ligne de démarcation claire entre un effort louable de rendre les services de l'État plus efficaces et un système qualifié par beaucoup de déshumanisant et attentatoire aux libertés

27 State v. Loomis, 881 N.W. 2d 749 (Wis. 2016), cert. denied, 137 S.Ct. 2290 (2017).
28 https://daily.jstor.org/what-happens-when-police-use-ai-to-predict-and-prevent-crime/
29 https://www.loc.gov/resource/rbpe.24404500/?st=text

individuelles. Le système chinois est donc un exemple à étudier pour tout gouvernement qui se lance dans un projet de transformation numérique de ses services.

Le 4e amendement de la Constitution limite l'utilisation de l'IA en tant qu'outil de surveillance

Outre le principe de *due process*, la deuxième limite constitutionnelle au déploiement de l'IA par l'État est la garantie contre des intrusions dans la vie privée fournie par le 4e amendement de la Constitution. Cet amendement interdit des intrusions du gouvernement dans le domicile privé sans autorisation judiciaire. Au fil du temps, la Cour suprême a étendu le domaine de protection accordé par ce 4e amendement pour couvrir non seulement le domicile, mais également des communications privées, l'intérieur de la voiture, et les déplacements en voiture. Cet amendement protège l'espace qu'un citoyen peut « raisonnablement s'attendre » à garder privé. Aujourd'hui, de nombreuses lois viennent le suppléer en encadrant les pouvoirs de la police, en exigeant par exemple l'obtention d'une réquisition judiciaire (*warrant*) pour accéder à des données privées. En matière d'espionnage, de terrorisme ou de sécurité nationale, les garanties de procédure sont moins fortes, mais elles existent. Jusqu'à présent, les tribunaux n'ont pas eu l'occasion de préciser comment le 4e amendement s'appliquerait aux techniques de surveillance facilitées par l'IA, mais la loi de 2020 de l'État de Washington sur la reconnaissance faciale[30] précise qu'une correspondance positive par ce système ne peut jamais être considérée comme une preuve suffisante, à elle seule, pour arrêter une personne ou fouiller son domicile. Il faut un faisceau d'autres indices pour atteindre le seuil de suspicion justifié (*probable cause*) exigé par la Constitution. De plus, chaque correspondance positive doit être vérifiée par un agent humain spécifiquement formé à l'utilisation de l'outil.

Conclusion prospective : enjeux d'avenir sous l'administration Trump

Comme nous l'avons vu, la régulation de l'IA aux États-Unis se situe surtout au niveau de chaque État fédéré. Certains États – la Californie, l'Illinois, l'État de Washington, l'État de New York – ont adopté des lois qui cherchent à protéger les citoyens contre les dérives de l'intelligence artificielle, et en premier lieu, contre les discriminations algorithmiques. Au niveau fédéral, les différentes agences, et en particulier la FTC, l'agence de protection des consommateurs, appliquent les lois existantes pour encadrer l'utilisation d'outils d'IA.

Quand bien même la nouvelle administration Trump et le Congrès pourraient essayer d'adopter une loi fédérale pour se substituer au *patchwork* de lois au niveau de chaque État, une telle loi ne serait pas susceptible de couvrir chaque cas d'usage, notamment sur les sujets qui demeurent sous la compétence de chaque État fédéré, comme la commande publique au niveau des administrations de ceux-ci, ou l'utilisation de l'IA par ces différentes administrations, y compris la police. Les États fédérés resteront donc des acteurs importants de la régulation de l'IA même sous la nouvelle administration Trump.

L'administration Trump pourrait avoir un impact important sur l'utilisation de l'IA au sein des administrations fédérales pour rendre les services de l'État plus efficaces. Une transformation numérique de l'État fédéral par l'utilisation de l'IA testerait les limites constitutionnelles qui protègent les citoyens contre les actions de l'État. En premier lieu, le principe de *due process* risquerait d'être mis à l'épreuve en cas d'automatisation des décisions administratives, que ce soit en matière de fiscalité, de justice pénale, d'immigration ou de gestion de ressources humaines au sein de l'État. Le principe de *due process* exige des procédures respectueuses des droits individuels, même si les contours exacts de celles-ci restent débattus.

Deuxièmement, la protection de la vie privée garantie par le 4e amendement de la Constitution limitera la possibilité pour l'État d'utiliser l'IA comme un outil de surveillance dans la lutte contre la criminalité. La police américaine utilise déjà des outils d'IA dans des enquêtes ou pour optimiser le déploiement des forces de police. Les romans et films de science-fiction (*Minority Report* pour n'en citer qu'un) nous rappellent que l'IA pourrait permettre d'aller beaucoup plus loin, par exemple *via* la mise en place d'un système de surveillance, de profilage et de prédiction concernant chaque citoyen. Un tel système nécessiterait le détricotage de nombreuses lois fédérales qui protègent la vie privée des citoyens à l'égard de l'État, détricotage en théorie possible si les Républicains contrôlent les deux chambres

30 https://app.leg.wa.gov/RCW/default.aspx?cite=43.386&full=true#43.386.010

du Congrès. Mais surtout, ces mesures seraient immédiatement contestées comme une violation du 4e amendement de la Constitution.

Enfin, remplacer les fonctions de l'État par des algorithmes peut affaiblir la démocratie dans son intégralité, facilitant l'émergence du techno-autoritarisme (Lafrance, 2024). Paul Nemitz a identifié cette menace dès 2018 (Nemitz, 2018), rappelant que la démocratie est inefficace à dessein. Les frottements, les contradictions, les ambiguïtés et les compromis sont des éléments inhérents et nécessaires à une démocratie saine. Vouloir les éliminer au nom de l'efficacité conduirait au fascisme (Nemitz, 2023).

Winston Maxwell est professeur de droit à Télécom Paris – Institut polytechnique de Paris, où il co-dirige le programme « Operational AI Ethics » de l'école.

Références bibliographiques

Beaudouin V. et Maxwell W. (2023),
« La prédiction du risque en justice pénale aux États-Unis : l'affaire ProPublica-Compas », *Réseaux : communication, technologie, société*, 240 (4), pp. 71-109. https://hal.science/hal-04455840

Berman E. (2018),
"A government of laws and not of machines", *Boston university law review*, vol. 98, p. 1277.

Brennan-Marquez K., Henderson S. (2019),
"Artificial Intelligence and Role-Reversible Judgment", 109 J. of Crim. L. & Criminology 137.

Chouldechova A. (2017),
"Fair prediction with disparate impact: A study of bias in recidivism prediction instruments", *Big data*, 5 (2), pp. 153-163.

Bertholet C. et Létourneau L. (2017),
Ubérisons l'État ! Avant que d'autres ne s'en chargent, Armand Colin.

Coglianese C., Lehr D. (2017),
"Regulating by robot: administrative decision making in the machine-learning era", *The Georgetown law journal*, vol. 105, nº 5, p. 1147.

Crawford K., Schultz J. (2019),
"AI Systems as State Actors", *Columbia Law Review*, vol. 119, nº 7, pp. 1941-1972.

Danziger S., Levav J. et Avnaim-Pesso L. (2011),
"Extraneous factors in judicial decisions", *Proceedings of the National Academy of Sciences*, 108 (17), pp. 6889-6892.

Heaven W.D. (2023),
"Geoffrey Hinton tells us why he's now scared of the tech he helped build", *MIT Technology Review*, 2 mai, https://www.technologyreview.com/2023/05/02/1072528/geoffrey-hinton-google-why-scared-ai/

Huq A. Z. (2020a),
"A right to a human decision", *Virginia Law Review*, vol. 106, p. 611 *sqq*.

Huq A. Z. (2020b),
"Constitutional rights in the machine learning state", *Cornell Law Review*, vol. 105, p. 1875 *sqq*.

Lafrance A. (2024),
"The Rise of Techno-Authoritarianism", *The Atlantic*, 30 janvier.

Livermore M. A. (2016),
"The perils of experimentation", *Yale Law Journal*, vol. 126, p. 636 *sqq*.

Maxwell W. (2022a),
Le contrôle humain des systèmes algorithmiques – Un regard critique sur l'exigence d'un « humain dans la boucle », Mémoire d'Habilitation à diriger des recherches, Faculté de droit de l'université Paris 1 Panthéon-Sorbonne. https://hal.science/tel-04010389/document

Maxwell W. (2022b),
« La régulation des algorithmes aux États-Unis : quelles leçons pour l'Europe ? », *in* Bertrand B. (dir.), *La politique européenne du numérique*, Bruxelles, Bruylant.

Maxwell W. (2023),
"Meaningful Human Control to Detect Algorithmic", *in* Castets-Renard C. & Eynard J. (eds.), *Artificial Intelligence Law : Between Sectoral Rules and Comprehensive Regime. Comparative Law Perspectives*, Bruylant. Texte disponible sur https://ssrn.com/abstract=4551063

Nemitz, P. (2023),
"Democracy through law The Transatlantic Reflection Group and its manifesto in defence of democracy and the rule of law in the age of artificial intelligence", *European Law Journal*, 29 (1).

Nemitz P. (2018),
"Constitutional Democracy and Technology in the age of Artificial Intelligence", *Philosophical Transactions of the Royal Society A*, 18 août, DOI 10.1098/RSTA.2018.0089, https://ssrn.com/abstract=3234336 ou http://dx.doi.org/10.2139/ssrn.3234336

Paige M. A., Amrein-Beardsley A. (2020),
"'Houston, we have a lawsuit' : a cautionary tale for the implementation of value-added models for high-stakes employment decisions", *Educational researcher*, vol. 49, n° 5, 2020, pp. 350-359.

Park, J. (2020),
"Your honor, AI", *Harvard International Review*, 41 (2), pp. 46-48.

Roesler S. M. (2014),
"Federalism and local environmental regulation", *University of California Davis Law Review*, 48, p. 1111 *sqq*.

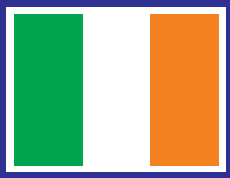
Sunstein C. R. (2022),
"Governing by Algorithm? No Noise and (Potentially) Less Bias", *Duke Law Journal*, vol. 71, n° 6, mars, pp. 1175-1205.

Werbach K. (2022),
"Orwell That Ends Well: Social Credit as Regulation for the Algorithmic Age", *University of Illinois Law Review*, vol. 2022, n° 4, septembre, pp. 1417-1475. https://ssrn.com/abstract=3589804 ou http://dx.doi.org/10.2139/ssrn.3589804

Whitt R. S. (2009),
"Adaptive policymaking: Evolving and applying emergent solutions for US communications policy", *Federal Communication Law Journal*, vol. 61, n° 3, article 2. https://www.repository.law.indiana.edu/fclj/vol61/iss3/2

Note réactive

Irlande



Publiée dans le cadre de l'activité de veille internationale en gestion publique du bureau de la recherche de l'IGPDE, la Note réactive décrit des expériences administratives réalisées dans d'autres pays du monde dans les cinq grands domaines du management public (budget et performance, gouvernance, relation à l'usager, emploi public, transformation numérique).

Irlande : former les agents publics à l'IA

Amicie de Tanoüarn

L'Irlande n'a pas attendu le lancement de ChatGPT en novembre 2022 pour réfléchir à la place de l'IA au sein de l'administration publique. Dès 2021, la stratégie nationale en matière d'IA, intitulée *AI Here for good*, est lancée par le département des dépenses publiques et de la réforme (DPER). Inscrite dans le cadre plus large de la stratégie d'innovation du pays, *Making innovation real*, et d'une volonté de renouveler la fonction publique irlandaise à l'horizon 2030, elle s'intéresse notamment aux questions de la formation, de la standardisation et de la certification des entreprises ainsi que des administrations pour permettre à l'Irlande de devenir l'un des *leaders* mondiaux de l'IA.[1]

Face à une technologie qui se répand progressivement dans tous les domaines de la société, relativement accessible à tout un chacun mais dont le développement comporte des risques éthiques, sociaux, économiques ou politiques, l'Irlande a rapidement fait émerger l'idée de former des agents publics afin qu'ils soient capables de comprendre les bases de l'intelligence artificielle, d'en porter un déploiement respectueux des valeurs publiques au sein de l'administration mais aussi d'en réguler les usages au cœur de la société. C'est dans cette optique que la Transformation Delivery Unit de la DPER lance le projet *AI foundation certificate for public servant*, un programme de formation ambitieux des agents publics aux subtilités de l'IA.

Former l'administration à l'IA, la stratégie pionnière de l'Irlande

Ce projet de formation est d'abord préfiguré avec l'aide d'acteurs nationaux, notamment l'Office of the government chief information officer (OGCIO), du responsable de la stratégie nationale en matière d'IA, d'acteurs publics supranationaux, notamment le programme *AI Accelerator* de la Commission européenne, d'institutions internationales, comme l'Observatoire de l'innovation dans le secteur public de l'OCDE, et d'experts reconnus, notamment le directeur du CeADAR, le centre irlandais de R&D appliquée à l'IA, afin de garantir l'alignement de son cadre et de son contenu sur les politiques et bonnes pratiques nationales et supranationales[2].

À l'issue de cette phase, le DPER crée, en 2021, l'*AI foundation certificate for public servant*, une formation certifiante destinée aux agents de l'administration irlandaise, organisée en partenariat avec l'université de Dublin. Cette formation poursuit deux objectifs : donner des compétences de base en IA et comprendre les conséquences de l'utilisation de l'IA, notamment au sein de l'administration publique[3]. S'inscrivant dans la politique du gouvernement irlandais d'amélioration des compétences de ses citoyens dans le domaine des technologies émergentes[4], cette formation est donnée sur 15 semaines avec 5 heures de cours en classe de 20 personnes. Mené par des intervenants experts dans certains domaines de l'IA, le programme couvre les principes fondamentaux de cette technologie, les cas d'utilisation dans le secteur public, les

1 https://enterprise.gov.ie/en/publications/publication-files/national-ai-strategy.pdf
2 https://oecd-opsi.org/innovations/ai-certification-ireland/
3 https://www.ernact.eu/NewsDetail.aspx?MediaNewsId=1564
4 https://www.gov.ie/en/publication/1e3e2-action/

implications éthiques et légales ainsi que les compétences pratiques nécessaires pour travailler avec des outils et données liés à l'IA.

Surtout, afin de permettre une mise en œuvre pratique des savoirs acquis lors de la formation, le DPER décide de sélectionner les candidats à la formation à partir de leur participation obligatoire à de potentiels projets d'IA ou des actions en cours de déploiement au sein de l'administration[5]. Conçue autour d'une participation active des formés, *via* notamment des travaux en groupe d'analyses de données ou de programmation, l'idée est donc bien celle de transmettre suffisamment de compétences techniques aux fonctionnaires afin qu'ils puissent piloter des projets IA au sein de l'administration et discuter d'égal à égal avec des prestataires extérieurs[6].

Enfin, le DPER organise une campagne de communication auprès de réseaux établis dans l'ensemble du secteur public irlandais, insistant aussi sur la prise en charge des frais de formation (environ 1 500 euros par participant), pour inciter les agents à se former à l'intelligence artificielle.

Un franc succès, restant néanmoins à évaluer

Depuis son lancement, la certification à l'intelligence artificielle pour les agents du service public irlandais a été délivrée à environ 1 300 agents. Ces derniers proviennent de services tels que la santé, l'éducation, l'administration centrale… et sont impliqués dans divers projets, par exemple d'évaluation de demandes frauduleuses, de création d'algorithmes d'extraction de données, d'évaluation de la quantité de lumière artificielle émise par l'Irlande ou de *chatbots* destinés à mieux orienter les parents d'enfants autistes vers les prestations auxquelles ils ont droit[7].

Le succès au sein de l'administration semble probant : d'une part les demandes de formation sont systématiquement excédentaires par rapport au nombre de places prévues[8] ; d'autre part, les premiers retours des participants à la formation montrent un taux de satisfaction générale à hauteur de 80 %, s'expliquant notamment par son ancrage dans la pratique et l'orientation des cours vers la mise en œuvre et la prise de décision pour des projets IA. Si aucune évaluation officielle de ce dispositif n'a encore eu lieu, en fin d'année 2024, une évaluation centrée sur son efficacité à long terme, l'amélioration des compétences en IA des agents et l'impact des projets IA accompagnés sur les politiques publiques qui les concernent, devait être menée par l'OGCIO, à partir d'entretiens réflexifs et d'analyses quantitatives[9]. Il est également prévu de mettre en œuvre des dispositifs de capitalisation des premières expériences de la formation, par exemple en mettant en place des webinaires accessibles à tous les agents, présentant les projets d'IA réussis et non réussis, ou en formant un vivier de talents et d'expériences, *via* la création de répertoires recensant les projets IA conduits et les professionnels certifiés.

Alors que la stratégie irlandaise, documentée dans la *Civil Service Renewal 2030 Strategy – Building on our Strengths* prévoit que 90 % des services publics seront fournis par voie numérique à l'horizon 2030, l'Irlande a donc pris au sérieux la nécessité de développer les compétences numériques de ses agents, ce qui pourrait être une source d'inspiration pour de nombreux autres pays.

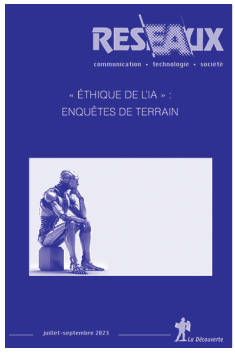
Amicie de Tanoüarn est apprentie « veille internationale et recherche en gestion publique comparée » au sein du bureau de la recherche de l'IGPDE.

5 https://www.interregeurope.eu/good-practices/ai-foundation-certificate-for-public-servants
6 https://oecd-opsi.org/innovations/ai-certification-ireland/
7 *Ibid.*
8 *Ibid.*
9 https://www.caidp.org/global-academic-network/ai-policy-clinic/

L'Œil du chercheur

L'Œil du chercheur propose une présentation synthétique de thèses récemment soutenues dans le champ de la gestion publique ainsi que des résumés d'articles marquants publiés dans un bouquet de revues spécialisées en sciences de gestion. Des informations d'ordre général peuvent également figurer sur les publications et les événements de l'IGPDE.

Revue des articles



« Éthique de l'IA » : enquêtes de terrain

Collectif

Numéro de la revue *Réseaux*, n° 240, 2023/4, éditions La Découverte, 284 pages, accessible en ligne : https://shs.cairn.info/revue-reseaux-2023-4?lang=fr

Le thème

Ce numéro spécial s'intéresse à un sous-champ de l'intelligence artificielle, l'éthique de l'IA (*AI ethics*). Partant du constat que les avancées exponentielles de l'IA depuis le milieu des années 2010 avaient conduit pour la première fois à la transformer en objet de controverses dans l'espace public, les auteurs s'intéressent aux manières dont les acteurs de l'IA, qu'ils soient publics ou privés, ont appréhendé les risques non techniques identifiés lors de ces controverses. Ils portent ainsi leur attention sur les interrogations qui ont émergé quant aux usages, aux conséquences sociales ou aux régulations juridiques de l'IA et sur les réponses qui y ont été apportées.

Les données

Ce numéro est composé de cinq articles ayant en commun de réunir des travaux empiriques et des enquêtes de terrain. Il s'appuie néanmoins sur une variété importante de méthodologies issues des sciences sociales : revue de la littérature grise, analyse quantitative de corpus (rapports administratifs, d'organisations internationales, de structures non gouvernementales, d'experts…), enquête par entretiens, étude de cas, observation ethnographique… Cette diversité méthodologique éclaire les enjeux structurant l'éthique de l'IA sur de multiples terrains : justice pénale, radiologie, entreprises de la *tech*, hôpitaux…

Les résultats

De nombreux résultats émergent de ces travaux. Le numéro montre que l'IA s'est progressivement imposée comme un enjeu public et révèle comment les différents acteurs se sont positionnés pour en appréhender les risques. S'il est aujourd'hui communément admis que certains principes éthiques sont nécessaires à la régulation des usages de l'IA (transparence, non-discrimination, explicabilité…), l'interprétation qui est faite de ces principes diffère pour plusieurs raisons. En effet, il est en premier lieu délicat de rendre techniquement opérationnelle la traduction de principe moraux. Cet enjeu est de plus accentué par le fait que l'éthique est parfois subordonnée à d'autres objectifs organisationnels (rentabilité, performance…) et que son interprétation peut diverger entre les acteurs impliqués sur le développement/l'usage d'un même outil IA, notamment en fonction de leurs métiers. À ce titre, les auteurs de ce numéro ont le mérite de démontrer que si de nombreux acteurs, notamment privés, cherchent des solutions techniques pour intégrer les enjeux éthiques, ces derniers ne peuvent être pris en considération sans une réflexion plus générale sur les conséquences organisationnelles des usages de l'IA.

Revue des articles

L'adoption de l'intelligence artificielle pour le développement de services publics intelligents

Aurélie Simard

Article paru dans *Communication, technologies et développement*, n° 11, « Intelligence artificielle et innovation sociale », sous la direction d'Alain Kiyindou, Étienne Damome et Noble Akam, 2022. Accessible en ligne : https://journals.openedition.org/ctd/6904

Le thème

Partant du constat que l'IA est aujourd'hui jugée par de nombreux acteurs publics comme suffisamment mature pour apporter de nombreux bénéfices à l'action publique, l'autrice de cet article se penche plus spécifiquement sur les facteurs d'adoption de l'IA pour mieux gérer la relation entre l'administration et les usagers des services publics. Elle étudie une solution expérimentale d'IA de gestion des contacts entre des conseillers Pôle emploi (dorénavant France Travail) et les demandeurs d'emploi, et analyse plus précisément un dispositif, les contacts *via* messages (CVM). Dans le contexte de l'accroissement des contacts dématérialisés entre l'administration et les usagers, l'introduction de l'IA dans les CVM poursuit un triple objectif de simplification du travail des conseillers : l'identification et l'authentification du demandeur d'emploi à l'origine de la demande ; la proposition d'une catégorie pour le courrier électronique ; la suggestion automatique d'une réponse ajustable par le conseiller.

Les données

Souhaitant à la fois décrire la perception des conseillers sur la valeur d'usage de l'IA et rendre compte de leur appréhension des risques de dérives, l'autrice a mené une triple phase d'échanges avec 15 collaborateurs de trois agences Pôle emploi, une en Île-de-France et deux en Occitanie. Elle s'appuie sur la méthodologie suivante : entretiens individuels semi-directifs, observation participante auprès des conseillers utilisant le dispositif et restitution auprès des conseillers des résultats intermédiaires obtenus, suivie d'une discussion. Les données collectées ont ensuite été retraitées grâce à une analyse thématique de contenus.

Les résultats

L'enquête aboutit à trois ensembles de résultats. Premièrement, elle soulève l'importance des éléments de contexte dans l'intégration d'un dispositif d'IA dans le travail administratif, en l'occurrence l'importance des autres canaux de contact (téléphone, accueil physique, numérisation des documents…) dont la mise en cohérence complexifie l'activité des conseillers. À cet égard, le gain de temps et l'amélioration des délais de réponse permis par l'IA apparaissent comme les objectifs les plus importants aux yeux des conseillers. Deuxièmement, la valeur d'usage perçue est importante, mais variable selon les trois fonctions attribuées à l'IA dans le cadre de cette expérimentation. Si l'identification automatique permet un gain de temps apprécié et si la catégorisation des messages facilite et augmente la coopération entre conseillers et équipes de direction pour identifier la meilleure expertise possible, les suggestions de réponse recueillent des avis plus nuancés car la standardisation induite nécessite fréquemment des ajustements en fonction des situations singulières des demandeurs d'emploi. Troisièmement, si les risques de discrimination et de profilage ont été considérés comme peu élevés, les risques de dérives justement liées à cette standardisation des réponses, ainsi que le risque de déshumaniser la relation avec les usagers, sont perçus comme importants.

Revue des articles

L'intelligence artificielle dans le secteur public : revue de littérature et programme de recherche

Marius Bertolucci

Article paru dans *Gestion et management public*, 2024/3, vol. 12, « Innovations stratégiques, organisationnelles et technologiques : analyse des défis et réponses possibles », et accessible en ligne : https://shs.cairn.info/revue-gestion-et-management-public-2024-3-page-71?lang=fr&tab=resume

Le thème

Cet article part du constat que l'IA devient un outil de plus en plus prisé des acteurs publics, par exemple en matière de police prédictive, de détection de la fraude ou de relation à l'usager, alors que, dans le même temps, certaines dérives ont mis en évidence les potentiels dangers de l'IA (discrimination, biais algorithmiques, violation de données…). Observant le développement de la recherche en management public sur le sujet, l'auteur se propose de réaliser une synthèse de la littérature universitaire existante afin d'analyser la manière dont ce champ de recherche est en train de se structurer.

Les données

L'auteur divise son étude en deux parties. La première, conçue à partir d'une analyse de trois revues systématiques de la littérature, examine la façon dont la discipline du management public se saisit actuellement de l'IA. La seconde partie propose une analyse quantitative et qualitative de 22 articles récents classés par les meilleures revues de management public, afin de faire émerger les applications, avantages et défis que l'IA présente pour l'action publique repérés dans cette littérature.

Les résultats

L'auteur met d'abord en évidence la nette progression des articles de recherche en management public consacrés à l'IA depuis 2020, laissant à penser qu'un champ de recherche est en voie de structuration aux côtés de champs universitaires historiquement associés à l'IA (informatique, médecine, sciences cognitives…). Cette montée en puissance des recherches en management public sur l'IA demeure malgré tout, selon l'auteur, trop embryonnaire et souffre d'un déficit de fondements théoriques et de recherches empiriques. L'article se termine par un agenda de recherche visant à stimuler l'exploration des usages de l'IA par l'action publique, l'auteur concluant sur un espoir de nouveaux travaux en la matière.

Revue des articles

Les trois grands défis posés par la gouvernance de l'intelligence artificielle et de la transformation numérique

Yannick Meneceur

Article publié dans *Éthique publique. Revue internationale d'éthique sociale et gouvernementale*, vol. 23, n° 2, « La gouvernance algorithmique », 2021, et accessible en ligne : https://journals.openedition.org/ethiquepublique/6323

Le thème

L'usage de l'intelligence artificielle et son introduction sur le marché sont contrôlés à la fois par le droit européen et le droit national. L'auteur montre toutefois que le manque de compréhension de la nature de l'intelligence artificielle comme du projet de société sous-tendu par cette technologie rend les contrôles et vérifications incomplets. Il prône trois actions susceptibles de compléter la régulation européenne : déconstruire le consensus sur la neutralité des technologies numériques et de l'intelligence artificielle ; objectiver les capacités des systèmes d'intelligence artificielle ; évaluer la soutenabilité environnementale de l'IA.

Les données

L'article s'appuie sur la mobilisation de différents travaux universitaires en IA, qu'ils proviennent des sciences sociales ou des sciences dures, ainsi que sur la littérature grise produite par certaines organisations internationales et nationales, notamment le ministère de la Transition écologique français ou le Conseil de l'Europe. Pour illustrer son propos, l'auteur présente des exemples concrets de dérives et de défaillances d'IA tout au long de l'article.

Les résultats

L'auteur de l'article plaide pour un usage proportionné de l'intelligence artificielle et des algorithmes décisionnels. Les impacts sociétaux et environnementaux de cette technologie étant élevés, seuls les secteurs où l'IA apporte une forte valeur ajoutée sociétale devraient l'utiliser. Dans les autres secteurs, c'est la décision humaine qui devrait être privilégiée. La situation précaire des travailleurs du clic, indispensables au fonctionnement des IA, la conjonction des biais humains et des biais algorithmiques, les conséquences de l'usage de l'IA sur les individus dans certaines organisations sont autant d'exemples mis en avant par l'auteur pour appeler à la prudence et à une utilisation raisonnée de l'IA.

Thèses et rapports

Intelligence artificielle et transformation de l'évaluation de programme

Steve Jacob, Seima Souissi, Loïc Duplantis

Rapport publié dans le cadre de la Chaire de recherche sur l'administration publique à l'ère numérique, université Laval, Québec, 2023, et accessible en ligne : https://www.administration-numerique.chaire.ulaval.ca/sites/administration-numerique.chaire.ulaval.ca/files/uploads/bureau/IA%20et%20%C3%A9valuation.pdf

Le thème

L'évaluation des politiques publiques nécessite la collecte d'un grand nombre de données complexes et hétérogènes, à analyser en un temps potentiellement restreint, et cela sur l'ensemble du cycle de vie d'une politique publique, de sa conception à sa mise en œuvre. Compte tenu du potentiel que présentent certaines technologies d'intelligence artificielle, comme le traitement automatique, le traitement du langage naturel, ou la vision par ordinateur, le rapport questionne leur usage pour repenser l'évaluation appliquée aux politiques publiques.

Les données

Certains travaux pionniers ont cherché à analyser les utilisations de l'IA en matière d'évaluation. Les auteurs réalisent une revue de littérature des différentes recherches exploratoires dans le domaine, et la restituent en identifiant les différents sous-domaines de l'évaluation (définition du périmètre d'évaluation, sélection et analyse des documents, étude des inférences causales, analyse des effets de politiques publiques…) Ils introduisent ensuite une dimension plus prospective pour discuter de la possibilité de faire émerger, grâce à l'IA, une évaluation continue dans toutes les phases du cycle de vie des politiques publiques.

Les résultats

Bien que les auteurs fassent le constat d'une faible application de l'intelligence artificielle à l'évaluation des politiques publiques, ils soulignent que les différentes expériences conduites montrent que son usage peut se répandre tout au long du processus d'évaluation. Les capacités de calcul de l'IA permettraient de réaliser une évaluation en continu de chaque politique publique, à la triple condition d'intégrer les avancées technologiques les plus récentes, de veiller à un usage responsable et éthique des données et d'assurer une transformation des métiers et des compétences.

Thèses et rapports

Promouvoir des modèles d'intelligence artificielle frugale pour et par les politiques publiques

Maxime Amisse, Mélissa Faur, Lucie Gonard, André Orcesi

Rapport de Groupe d'analyse de l'action publique réalisé dans le cadre du Mastère « Politiques et action publiques pour le développement durable » (PAPDD), École des Ponts Paris Tech, 2024.
https://hal.science/hal-04510171v1/document

Le thème

Les investissements dans le domaine de l'IA en France et dans le monde ont augmenté de façon exponentielle depuis quelques années. Toutefois, afin d'atteindre des résultats satisfaisants, l'intelligence artificielle a besoin d'une quantité massive de données et de calculs de plus en plus complexes, opérés à partir d'infrastructures de plus en plus puissantes. Ce fonctionnement général de l'IA a un poids environnemental important qui met en tension la performance, les gains permis par l'IA et leur coût environnemental. Ce travail de recherche investit les différentes définitions d'une IA frugale, notamment en termes d'optimisation des quantités de données utilisées, d'architecture des algorithmes, de choix de matériel, ou de source d'énergie utilisée. Sont examinées dans ce rapport les manières dont l'administration s'approprie ce concept mais aussi le rôle que la recherche publique peut tenir pour participer à l'intégration de l'IA frugale dans nos usages.

Les données

La recherche conduite s'appuie sur des entretiens avec des acteurs du domaine de l'IA frugale et une analyse bibliographique des publications, des appels à projets et de la littérature grise (rapports de l'observatoire Data publica, rapport Villani, rapports de l'OCDE…). Quatre cas d'étude sont par ailleurs approfondis : l'Agence de Services et de Paiement – le plus important organisme payeur européen –, la collectivité locale de Noisy-le-Grand et deux établissements publics administratifs, Météo France et l'Institut national de l'information géographique et forestière.

Les résultats

Les recherches aboutissent à un constat nuancé sur le déploiement de l'IA frugale. D'une part, elles montrent une démarche proactive pour valoriser une IA frugale au sein de l'État, notamment au sein de l'administration (Ecolab aux ministères Aménagement du territoire Transition écologique, Etalab au sein de la direction interministérielle du numérique) et, plus généralement, au sein d'un écosystème varié (PME, associations, *start-up*, communauté scientifique…) doté de certains moyens financiers, humains et infrastructurels. D'autre part, l'IA frugale semble regrouper un ensemble d'interprétations assez différentes, se matérialisant dans des actions proches du *greenwashing*, ce qui rend difficile une normalisation consensuelle. Enfin, l'arbitrage entre performance économique et performance environnementale penche bien souvent pour la première.

Thèses et rapports

La justice algorithmique en chantier : sociologie du travail et des infrastructures de l'intelligence artificielle

Camille Girard-Chanudet

Thèse de doctorat en sociologie sous la direction de Nicolas Dodier et Valérie Beaudouin, soutenue le 4 décembre 2023 à l'EHESS.
https://theses.fr/2023EHES0141

Autour de 2016 apparaissent en France les premiers algorithmes d'apprentissage automatique orientés vers le traitement des décisions de justice mises en *open data*. À partir d'une enquête empirique alliant ethnographies, entretiens et analyse de matériau documentaire, cette thèse entre dans les coulisses du chantier de l'intelligence artificielle (IA) juridique, pour mieux analyser les ressorts et implications sociales du développement de ces outils fortement controversés. Elle suit d'abord la trajectoire de ces objets techniques, ainsi que des représentations et tensions qui les accompagnent, depuis le monde de l'entrepreneuriat numérique jusqu'aux portes des tribunaux ; elle met en évidence, ce faisant, les requalifications successives du « concept-frontière » d'IA, qui agrège autour de lui, en fonction des sens et des finalités qui y sont associés, des configurations plurielles d'acteurs. La thèse plonge ensuite dans les infrastructures informationnelles qui constituent le cœur du dispositif de l'IA juridique : elle s'intéresse aux données qui servent de fondement aux algorithmes d'apprentissage automatique et aux processus de requalification, centralisation, formatage et mise en circulation auxquels elles sont soumises pour permettre l'entraînement des modèles. Elle documente enfin le travail de conception l'IA, à partir d'une ethnographie conduite au sein d'une équipe pluridisciplinaire constituée à la Cour de cassation : elle montre la façon dont les activités de traduction, d'articulation, de négociation et les choix opérés par des acteurs aux expertises hétérogènes façonnent progressivement le dispositif algorithmique. À chaque niveau, cette recherche met en évidence la dimension intrinsèquement sociale de l'IA. Elle souligne les processus d'hybridation d'expertises techno-juridiques se produisant au contact de l'IA, qui contribuent à renforcer des lignes de fracture au sein des mondes traditionnels du droit, entre élites et professionnels de terrain.

Thèses et rapports

L'adoption de l'intelligence artificielle par le chef d'établissement : l'aide à la décision algorithmique pour organiser le temps scolaire

Nathalie Glais

Thèse de doctorat en sciences de l'éducation sous la direction de Béatrice Mabilon-Bonfils, soutenue le 15 décembre 2023 à CY Cergy Paris Université.
https://theses.fr/2023CYUN1221

Cette thèse exploratoire appréhende le déploiement de l'intelligence artificielle en milieu scolaire, dans le contexte de la stratégie nationale de l'intelligence artificielle de la France et des recommandations de l'UNESCO en la matière. Elle est centrée sur les chefs d'établissement du secondaire en France qui utilisent l'IA comme outil d'aide à la décision pour réaliser les emplois du temps scolaires. Elle investit plus spécifiquement la relation qui se noue entre l'humain et l'artefact et la possibilité d'une appropriation du logiciel par les chefs d'établissement qu'ils soient en situation d'adoption consentie ou imposée.. Sur le plan méthodologique, la thèse analyse le mode d'emploi prescriptif du logiciel d'IA à l'aide de l'analyse des tâches (TMTA), puis le confronte à la pratique professionnelle de chefs d'établissement saisie par des entretiens semi-directifs. Enfin, pour vérifier la solidité du modèle utilisé, 916 réponses à un questionnaire distribué aux utilisateurs du logiciel d'emploi du temps sont analysées statistiquement. Les résultats de la thèse montrent notamment que le logiciel d'IA utilisé par les chefs d'établissement apparaît comme rigide et centré sur le savoir des ingénieurs, ce qui ne le rend pas complètement idoine aux attentes des utilisateurs, alors même que ceux-ci sont en demande d'une aide technologique performante. Ils rappellent donc la nécessité d'adapter la technologie aux besoins métiers.

Thèses et rapports

Smart city, du concept à l'expérience quotidienne : l'exemple de Songdo en Corée du Sud

Suzanne Peyrard

Thèse de doctorat en géographie sous la direction de Valérie Gelézeau soutenue le 5 décembre 2023 à l'EHESS.
https://theses.fr/2023EHES0145

Depuis la fin des années 2000, la ville intelligente (*smart city*) s'impose comme une panacée pour les urbanistes, les promoteurs et les acteurs publics. Sans que ses fondements théoriques ni son développement pratique ne soient définis, ce modèle urbain circule à l'échelle internationale en même temps qu'un certain nombre d'équipements « intelligents » dont le smartphone fait partie. À travers une enquête de terrain (2018-2022) menée à Songdo, une *smart city* en cours de construction en Corée du Sud, cette thèse interroge notre manière d'habiter, c'est-à-dire de penser, théoriser, construire et pratiquer un espace urbain de plus en plus façonné par les technologies de l'information et de la communication (TIC). Elle met d'abord en perspective le discours des habitants avec celui des « sponsors urbains » (les développeurs qui influencent la gouvernance urbaine), exposant les écarts inévitables entre la réalisation d'un mégaprojet urbain pris dans une mécanique financière capitaliste et la réalité quotidienne d'habitants souvent désenchantés face à l'attente de concrétisation. Avant d'être qualifiée de *smart*, Songdo apparaît comme une nouvelle ville de la région de Séoul conçue pour une élite sociale en quête d'une modernité symbolisée par des complexes d'appartements (*apateu danji*). Une enquête ethnographique d'un laboratoire d'ingénieurs permet ensuite d'analyser leurs actions d'informatisation urbaine et de reconsidérer de manière critique dans quelle mesure la standardisation de la production urbaine impacte l'informatisation de la ville, allant jusqu'à normaliser l'intelligence artificielle (IA). Malgré un développement technologique fondé sur l'IA, donc estimé plus performant que d'autres mégaprojets urbains par ses concepteurs, les résultats de l'enquête révèlent que l'expérience urbaine de Songdo est comparable à celle vécue dans d'autres villes sud-coréennes où l'utilisation d'un smartphone pour naviguer dans l'espace urbain est devenue un « fait social total ». En définitive, Songdo constitue un terrain d'étude exemplaire d'exploration de doctrines et d'idéologies sous-jacentes à une production urbaine standard du XXIe siècle.

Thèses et rapports

Assemblée nationale (2024), ***Rapport d'information sur les défis de l'intelligence artificielle en matière de protection des données personnelles et d'utilisation du contenu généré.***

https://www.assemblee-nationale.fr/dyn/16/rapports/cion_lois/l16b2207_rapport-information

Ce rapport se penche sur les implications de l'IA générative en termes de protection des données personnelles et de gestion du contenu généré. Il met en lumière les défis réglementaires et éthiques induits, soulignant la nécessité de mettre à jour les cadres législatifs pour encadrer efficacement cette technologie émergente. Il propose notamment une vérification préalable qui comprend la certification et l'identification des contenus, une réglementation des plaintes et un système de punition, ainsi que le déploiement des compétences spécialisées.

Conseil d'État (2022), ***Intelligence artificielle et action publique : construire la confiance, servir la performance.***

https://www.conseil-etat.fr/publications-colloques/etudes/intelligence-artificielle-et-action-publique-construire-la-confiance-servir-la-performance

En affirmant que l'utilisation de l'IA pourrait améliorer la qualité du service public rendu aux citoyens, le Conseil d'État plaide pour une stratégie de l'IA au service de la performance publique. Il promeut la mise en œuvre de lignes directrices permettant un déploiement progressif de l'intelligence artificielle dans les services publics pour répondre aux besoins des Français et établit sept principes pour une IA de confiance : la primauté humaine, la performance, l'équité et la non-discrimination, la transparence, la sûreté (cybersécurité), la soutenabilité environnementale et l'autonomie stratégique.

Cour des comptes (2023), ***La stratégie nationale de recherche en intelligence artificielle.***

https://www.ccomptes.fr/fr/publications/la-strategie-nationale-de-recherche-en-intelligence-artificielle

Ce rapport cherche à évaluer si la stratégie nationale de recherche a permis de renforcer la position de la France aux niveaux mondial et européen, si elle a accéléré la structuration de l'écosystème français en IA, si, en matière de recherche, elle a créé des pôles d'excellence efficaces et efficients et si elle a amélioré la prise en compte des enjeux éthiques.

Cour des comptes (2024), ***L'intelligence artificielle dans les politiques publiques : l'exemple du ministère de l'Économie et des Finances.***

https://ccomptes.fr/fr/publications/lintelligence-artificielle-dans-les-politiques-publiques-lexemple-du-ministere-de

Ce rapport analyse le déploiement de l'IA au sein des ministères économiques et financiers. En se penchant sur les trente-cinq programmes d'intelligence artificielle qui y sont déployés depuis 2015, la

Cour met en évidence une bonne maîtrise des enjeux technologiques, mais une moindre vigilance sur les enjeux éthiques, RH ou environnementaux. Elle en appelle ainsi à un pilotage ministériel robuste, à une meilleure évaluation des gains de productivité générés et à une anticipation et un partage d'informations plus développés.

Défenseur des Droits (2024), *Algorithmes, systèmes d'IA et services publics : quels droits pour les usagers ? Points de vigilance et recommandations.*

https://www.defenseurdesdroits.fr/algorithmes-intelligence-artificielle-et-services-publics-2024

Face au nombre croissant de décisions administratives individuelles prises à partir de résultats livrés par des algorithmes ou systèmes d'IA, ce rapport s'inquiète des risques qu'induit cette « algorithmisation » des services publics pour les droits des usagers. Il présente plusieurs recommandations afin que les garanties prévues par la loi soient pleinement concrétisées.

Premier ministre (2024), *IA : notre ambition pour la France.*

https://www.economie.gouv.fr/files/files/directions_services/cge/commission-IA.pdf?v=1717773778

Ce rapport s'intéresse aux potentialités de croissance et aux risques de l'IA, relativisant notamment son impact sur le chômage et rappelant les enjeux de compétitivité, d'indépendance et de progrès qu'elle porte. Il propose dès lors six actions, dont un plan de sensibilisation et de formation de la nation, la réorientation de l'épargne vers l'innovation et la création d'un fonds dédié à l'IA doté de 10 milliards d'euros, des investissements spécialisés dans la puissance de calcul ou une gouvernance mondiale de l'IA.

Sénat (2020), *Rapport d'information sur l'empreinte environnementale du numérique.*

https://www.senat.fr/rap/r19-555/r19-555_mono.html

En démontrant que le numérique a longtemps été l'angle mort des politiques environnementales, ce rapport réalise une évaluation inédite de l'empreinte carbone du numérique en France et trace en conséquence une feuille de route poursuivant les objectifs suivants : faire prendre conscience aux utilisateurs du numérique de son impact environnemental, limiter l'empreinte carbone du numérique, développer des usages du numérique écologiquement vertueux et des infrastructures moins énergivores.

Cette rubrique a été préparée par **Amicie de Tanoüarn**, apprentie « veille internationale et recherche en gestion publique comparée » au sein du bureau de la recherche de l'IGPDE, et **Edoardo Ferlazzo**, chef du département Gestion publique comparée.

IA : Intelligence de l'Action publique ?

L'IGPDE a mis en ligne son nouveau ***motion design***, présenté publiquement le 14 novembre 2024 lors des **Rencontres internationales de la gestion publique**, qui portaient sur le thème : « Gouverner (par) l'IA : l'action publique à la croisée des chemins ».

https://urlr.me/xU7NRF

YouTube

Après avoir retracé l'histoire de l'intelligence artificielle, de ses origines dans les années 1940 à ses déploiements plus contemporains dans des champs de recherche variés (IA connexionniste, IA symbolique, *deep learning*…), cette courte vidéo, illustrée par des exemples d'utilisation de l'IA dans plusieurs pays européens, aborde les manières dont l'IA pourrait être déployée au sein de l'action publique. En présentant les opportunités que les avancées technologiques les plus récentes offrent pour repenser l'action publique, elle évoque également les grands enjeux de l'IA (souveraineté, cybersécurité, transparence, éthique…) pour la puissance publique.

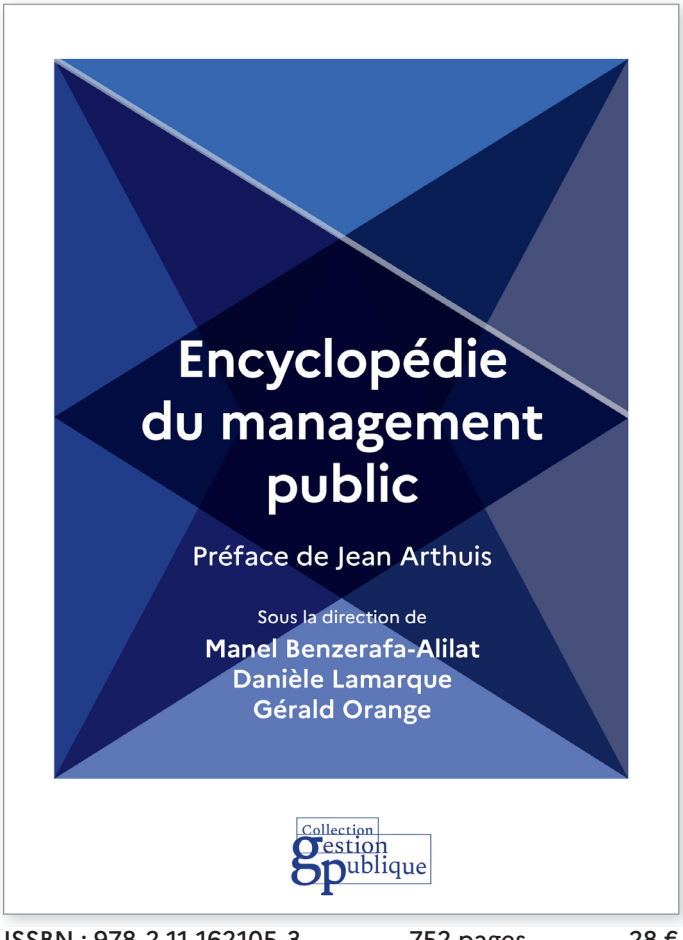
Ce *motion design* a été réalisé sur la base de travaux académiques. Une **bibliographie** et une **revue de littérature** rédigées par Edoardo Ferlazzo, chef du département de gestion publique comparée au sein du bureau de la Recherche de l'IGPDE, sont accessibles à l'adresse https://www.economie.gouv.fr/igpde-editions-publications/ia-action-publique

SOMMAIRE DU NUMÉRO PRÉCÉDENT

N° 22 (2024/3) – CONCILIER LES TEMPS DE L'ACTION PUBLIQUE

INSTITUT DE LA GESTION PUBLIQUE ET DU DÉVELOPPEMENT ÉCONOMIQUE

Gestion publique



ISSBN : 978-2-11-162105-3 752 pages 28 €

https://www.economie.gouv.fr/igpde-editions-publications
Pour se procurer l'ouvrage auprès de l'éditeur :
recherche.igpde@finances.gouv.fr

Édition numérique accessible sur
https://books.openedition.org/igpde/15291

OpenEdition Books

Action publique. Recherche et pratiques est une revue trimestrielle de l'Institut de la gestion publique et du développement économique qui s'adresse aux acteurs publics, aux universitaires et aux étudiants désireux de s'informer sur l'évolution des savoirs dans le champ de l'action publique. Cette revue fait intervenir chercheurs et praticiens du secteur public à travers des contributions sous forme d'articles, d'entretiens et de notes d'analyses décrivant des expériences administratives dans divers pays. La revue présente également des thèses récentes et des articles de recherche marquants en gestion publique.

CAIRN.INFO
MATIÈRES À RÉFLEXION