STEVE JACOB
SÉBASTIEN BROUSSEAU

QUAND L'ALGORITHME DÉCIDE

L'État, L'IA et nous

Préface du protecteur du citoyen
Marc-André Dowd

PUL

QUAND L'ALGORITHME DÉCIDE

L'État, l'IA et nous

ÉTHIQUE, IA ET SOCIÉTÉ

Collection dirigée par Lyse Langlois

Les innovations technologiques en intelligence artificielle sont un vecteur important du développement économique et peuvent contribuer significativement à l'amélioration des conditions de vie aujourd'hui et pour les générations futures. Leur impact soulève toutefois des enjeux majeurs, en matière d'emploi, de santé, d'éducation, de sécurité, de démocratie, de justice ou encore d'éthique. Ces enjeux traversent les frontières, nécessitent une réflexion et des actions pour lesquelles toute la communauté de recherche est appelée à jouer un rôle clé, et ce, en étroite collaboration avec les parties prenantes des milieux de l'industrie, du gouvernement et de la société civile. Cette collection présente des ouvrages érudits mais également des ouvrages de vulgarisation scientifique qui portent plus particulièrement sur les impacts sociétaux et éthiques de l'intelligence artificielle et du numérique.

Série
Points de vue

La série « **Points de vue** » se veut une source d'information incontournable pour explorer et comprendre les impacts sociétaux de l'intelligence artificielle et les données qui sont nécessaires au fonctionnement des modèles algorithmiques. Elle entend démystifier la recherche tout en préservant son intégrité intellectuelle. Elle met l'accent sur la diffusion et le partage des connaissances en offrant des points de vue éclairés et accessibles qui clarifient les complexités technologiques émergentes et la manière dont ces avancées influencent nos vies quotidiennes, nos institutions et nos interactions sociales. Chaque ouvrage agissant comme une passerelle entre le monde universitaire et le grand public, **Points de vue** aspire à sensibiliser sur les enjeux éthiques et sociaux de même que sur les défis et opportunités liés à l'IA en adoptant une approche pédagogique.

Une liste des autres titres parus dans la collection apparaît à la fin de l'ouvrage.

STEVE JACOB
SÉBASTIEN BROUSSEAU

QUAND L'ALGORITHME DÉCIDE

L'État, l'IA et nous

Préface du protecteur du citoyen
MARC-ANDRÉ DOWD

Presses de
l'Université Laval

Financé par le gouvernement du Canada
Funded by the Government of Canada

Canada

Nous remercions le Conseil des arts du Canada de son soutien.
We acknowledge the support of the Canada Council for the Arts.

Conseil des arts du Canada
Canada Council for the Arts

Les Presses de l'Université Laval reçoivent chaque année de la Société de développement des entreprises culturelles du Québec une aide financière pour l'ensemble de leur programme de publication.

SODEC Québec

Révision linguistique : Solange Deschênes
Mise en pages : Diane Trottier
Maquette de couverture : Laurie Patry

Imprimé au Canada

Dépôt légal 2e trimestre 2025
ISBN : 978-2-7663-0724-1
ISBN PDF : 9782766307258
ISBN ePub : 9782766307265

Les Presses de l'Université Laval
www.pulaval.com

PRÉFACE

Marc-André Dowd
Protecteur du citoyen

Le rôle d'un ombudsman, tel que le protecteur du citoyen, est de s'assurer qu'en toutes circonstances l'administration publique rend des décisions fondées sur le droit, justes, sans arbitraire ni discrimination. Il faut aussi s'assurer que l'équité est prise en considération et que les situations exceptionnelles qui sont susceptibles de se présenter vont trouver une réponse acceptable – et humaine – de la part de l'Administration, quelquefois même au-delà des petites cases prévues sur le formulaire.

C'est pourquoi le sujet de l'utilisation de l'intelligence artificielle dans la prestation des services publics suscite, dans la communauté des ombudsmans et des médiateurs, autant de fascination que d'inquiétude. D'une part, il faut reconnaître que les systèmes d'intelligence artificielle (SIA), utilisés comme outils d'aide à la décision par des agents de l'État, ont le potentiel évident de générer de l'efficience, de la cohérence dans les décisions et une certaine forme d'équité *formelle* dans la prise en considération des mêmes facteurs pour tous. D'autre part, ils inquiètent quant à leur capacité à prendre en compte des éléments, quelquefois intangibles, en présence de situations extraordinaires où la stricte application de la norme donnerait lieu à un résultat qui répugne au bon sens ou qui nous apparaîtrait… inhumain.

De même, l'essence du travail d'un ombudsman se situe dans le dialogue avec les agents de l'État concernant leurs décisions, de manière à s'assurer qu'elles prennent en considération l'ensemble des circonstances pertinentes dans l'évaluation de la situation d'un citoyen. Or, qu'en est-il lorsque d'opaques outils d'aide à la décision fondés sur l'IA sont utilisés et que l'on cherche, parmi les agents de l'État, la personne qui pourra nous expliquer quels facteurs sont pris en considération et avec quelle pondération ? En ce sens, le concept de *boîte noire*, inhérent à de tels systèmes, n'a rien pour rassurer. Et quand on commence à lire sur les mécanismes de l'apprentissage profond (*deep learning*), le vertige nous prend.

On saisit alors toute l'importance du problème des biais algorithmiques. Et c'est pourquoi j'ai trouvé la lecture de cet ouvrage de Steve Jacob et Sébastien Brousseau particulièrement éclairante. D'abord, les auteurs font très bien ressortir, avec l'aide d'exemples concrets, les problèmes majeurs qui risquent de survenir si l'on néglige de s'intéresser à cet aspect. Il s'agit ici de décisions de l'État rendues en violation de droits fondamentaux des citoyens, au premier chef, le droit à l'égalité. La mise à jour de biais sexistes ou racistes, introduits dans certains SIA, en constitue des exemples frappants.

Malgré le caractère aride de la matière, les auteurs démontrent avec beaucoup de pédagogie les biais qui sont susceptibles de se manifester, à toutes les étapes de production d'un SIA. On comprend aisément qu'il faut demeurer vigilant à tout moment, dans la définition du problème à résoudre, la collecte de données à utiliser, la modélisation, la validation et jusqu'au déploiement de la solution. La typologie des biais proposée par les auteurs est claire et elle

nous fait prendre conscience du caractère multiforme et un peu diffus du problème des biais algorithmiques.

La lecture de cet ouvrage pourrait ainsi s'avérer angoissante sans la présence du dernier chapitre. Les auteurs en appellent à réaffirmer avec force le principe selon lequel la prise de décision, même assistée par l'intelligence artificielle, doit toujours demeurer sous la responsabilité du personnel de l'État. De même, sans égard aux avancées technologiques, les organisations publiques demeurent soumises au droit. Ces deux éléments m'apparaissent incontournables.

Steve Jacob et Sébastien Brousseau nous invitent ensuite à envisager le développement de SIA en fonction de certaines balises : la pertinence (voire la nécessité) d'un tel système dans un contexte donné, l'acceptabilité sociale, les moyens à mettre en place pour garantir des données de qualité (en matière de disponibilité, de préparation adéquate et de représentativité), ainsi que les mesures à prendre pour assurer la transparence nécessaire, auprès de la population, dans l'utilisation des SIA par l'État.

Les auteurs proposent ainsi une véritable grille de réflexion sur l'utilisation des algorithmes dans le secteur public, alliant les diverses dimensions techniques et humaines pertinentes. Cet outil me sera précieux dans l'examen des situations qui ne manqueront pas de se présenter dans un proche avenir. J'invite aussi mes collègues gestionnaires publics à s'y référer dans tout projet de développement d'un système fondé sur l'intelligence artificielle. La démarche proposée permettra, sans contredit, de préserver nos valeurs démocratiques et de servir l'intérêt public.

TABLE DES MATIÈRES

LISTE DES ENCADRÉS ET DES FIGURES

REMERCIEMENTS

Cet ouvrage est le fruit d'une étude réalisée dans le cadre des activités scientifiques de l'Observatoire international sur les impacts sociétaux de l'intelligence artificielle et du numérique (Obvia) et de la Chaire sur l'administration publique à l'ère numérique de l'Université Laval. Cette recherche a bénéficié du soutien financier du Fonds de recherche du Québec et du Secrétariat du Conseil du trésor du gouvernement du Québec, que nous remercions pour leur appui.

L'aide précieuse de monsieur Richard Dufour, bibliothécaire-conseil à l'Université Laval, a été déterminante pour concevoir la stratégie de recherche documentaire afin d'accéder à des publications difficilement disponibles. Nous souhaitons également remercier Seima Souissi, Loïc Duplantis et Lissage Pascal pour leur travail de recherche documentaire et d'analyse de la littérature.

Certains passages de cet ouvrage ont été retravaillés et reformulés à l'aide d'outils d'intelligence artificielle tels que ChatGPT, Copilot ou DeepL. Les suggestions générées par ces outils d'intelligence artificielle ont été vérifiées et validées par les auteurs.

INTRODUCTION

Cet algorithme pourrait vous gâcher la vie !

Nous sommes souvent irrités d'entendre une voix automatisée répéter des phrases telles que « Votre appel est important pour nous » ou « Ce temps d'attente est bien involontaire de notre part » lorsque nous appelons une administration publique pour obtenir une information ou un service[1]. Pour le moment, aucune administration n'utilise des avertissements plus menaçants, tels que : « Attention, cet algorithme pourrait gâcher votre vie ! » ou « Les biais des processus automatisés qui vous ciblent ne sont pas de notre responsabilité ».

Avec l'utilisation accrue de l'intelligence artificielle (IA) dans le secteur public, ces avertissements pourraient toutefois devenir nécessaires. En effet, à l'échelle internationale, nous observons que, malgré les avantages que présentent les algorithmes d'aide à la décision dans le secteur public, ceux-ci produisent également des erreurs et alimentent la discrimination dans de nombreux pays. Le gouvernement des Pays-Bas a même dû démissionner, en 2021, à la suite d'un scandale impliquant son administration fiscale. Pour détecter les fraudes aux allocations familiales, l'administration fiscale

1. Le titre de l'introduction est emprunté à celui d'un dossier thématique de la revue *Wired* consacré à la discrimination algorithmique publié en mars 2023.

néerlandaise utilisait un algorithme qui attribuait un score de risque de fraude plus élevé aux ressortissants étrangers. Dans un rapport analysant ce scandale, intitulé *Les machines xénophobes*, Amnistie internationale dénonce l'automatisation croissante de la prestation de services publics et plaide pour une interdiction de ces « algorithmes racistes » (Amnesty International, 2021).

De nos jours, l'intelligence artificielle (IA) est utilisée par un nombre croissant d'organisations dans le monde, y compris au sein du secteur public. En recourant à l'IA, les organisations publiques aspirent à optimiser la rapidité, l'efficacité, l'objectivité et la qualité des processus, des services et des biens fournis à la population. L'intégration et le déploiement de systèmes décisionnels automatisés dans le secteur public s'accélérant considérablement, il est nécessaire d'examiner les implications de l'utilisation de l'IA par les organisations publiques, notamment en ce qui concerne les risques du renforcement des asymétries de pouvoir entre l'État et la population et les risques de préjudices que pourraient subir les personnes et les groupes marginalisés (Kuziemski et Misuraca, 2020).

En effet, des processus administratifs ou des décisions affectant la vie des citoyennes et des citoyens sont de plus en plus souvent guidés par des algorithmes d'IA. Ainsi, des diagnostics médicaux intègrent des modèles prédictifs établis à partir de vastes ensembles de données ; l'octroi de microprêts dans le cadre de programmes d'aide internationale s'appuie sur des jugements algorithmiques de solvabilité ; les décisions d'envoyer des travailleurs sociaux enquêter sur des cas potentiels de maltraitance envers les enfants sont guidées par des scores de risque basés sur des algorithmes. Les exemples se multiplient et de nouvelles illustrations de l'utilisation de l'IA dans le secteur public apparaissent chaque jour.

Parallèlement, une prise de conscience croissante émerge en ce qui concerne les répercussions engendrées par les biais de ces algorithmes. Les algorithmes de reconnaissance faciale qui fonctionnent moins bien pour des personnes aux traits féminins ou à la peau plus foncée (et encore moins pour ceux qui cumulent les deux caractéristiques) empêchent des personnes d'avoir accès à des services auxquels elles ont droit ; des modèles de prédiction de récidive qui classent des personnes de couleur comme présentant un risque significativement plus élevé que des individus blancs conduisent à des incarcérations injustes. Dans chaque cas, l'utilisation d'algorithmes risque de reproduire, voire d'aggraver, les inégalités sociales existantes (Fazelpour et Danks, 2021).

De plus en plus de chercheurs et d'observateurs expriment ainsi leurs préoccupations, non seulement face aux erreurs des systèmes d'IA dues à une mauvaise conception ou à des erreurs de programmation, mais aussi face à des risques, à long terme, plus sérieux, tels que la reproduction et l'ancrage, dans ces systèmes d'IA, de préjugés et de structures de pouvoir existants. D'un autre côté, des questions sont aussi soulevées quant à la capacité des systèmes algorithmiques à pouvoir s'adapter à des situations imprévues ou à exercer le degré de discernement dont l'humain est capable face à des situations sociales complexes et nuancées (Bannister et Connolly, 2020).

Le phénomène des biais algorithmiques a pris davantage de place dans le débat public avec le lancement d'applications d'IA générative facilement accessibles à tout le monde. Le lancement de ChatGPT d'OpenAI, en novembre 2022, a été qualifié de tsunami technologique et une myriade de systèmes concurrents ont rapidement vu le jour ; c'est ainsi que les Claude (Anthropic), Gemini (Google), Copilot (Microsoft), Pi (Inflection AI) et LLaMA (Meta), pour n'en

nommer que quelques-uns, font maintenant partie de notre environnement et suscitent la fascination et l'inquiétude. L'intérêt croissant pour ces types d'outils, attribuable à leur interface interactive axée sur le langage naturel tel que nous le parlons couramment, a mis en lumière tant les potentialités que les limites inhérentes à l'intelligence artificielle en général et à ces grands modèles de langues génératifs (*large language models*[2]) en particulier. La nature intrinsèquement interactive de ces modèles de langues génératifs a offert une occasion unique, tant aux experts du domaine qu'aux novices, d'évaluer la qualité et les imperfections des réponses ou des travaux produits par ces applications d'IA. Bien que de nombreux observateurs s'accordent sur le potentiel d'évolution rapide de cette technologie, les erreurs, imprécisions, affabulations, hallucinations et biais dans ses résultats suscitent de nombreuses préoccupations (Nkonde, 2023 ; Van Voorhis, 2023).

Ces réactions sont légitimes et nécessitent des réponses adéquates. Toutefois, l'intérêt récent pour les prouesses et les défaillances des grands modèles de langues génératifs ne devrait pas éclipser les problèmes, associés aux biais algorithmiques, causés par d'autres applications d'IA moins interactives ou anthropomorphes, telles que les systèmes prédictifs de criminalité, les algorithmes de recommandation ou les logiciels de reconnaissance faciale qui utilisent des algorithmes d'apprentissage automatique (*machine learning*). Ces technologies d'IA procèdent souvent de manière invisible pour le grand public, car elles fonctionnent en arrière-plan

2. Les termes techniques provenant de l'anglais sont bien souvent traduits de différentes manières en français. Afin d'uniformiser la traduction, nous nous référons à la traduction fournie par le *Grand lexique français de l'intelligence artificielle* disponible sur le site DataFranca.org. Ce site contient de nombreuses informations permettant de comprendre le fonctionnement de l'IA.

et impliquent rarement des interactions directes. En dépit de cette discrétion, les algorithmes d'apprentissage automatique soulèvent également des problèmes et des défis complexes. Malgré les avancées notables dans la compréhension de ces enjeux, une grande partie du travail reste à faire, en particulier en ce qui concerne la sensibilisation des concepteurs et des utilisateurs d'outils technologiques intégrant l'IA.

Les gouvernements sont confrontés à l'ensemble de ces enjeux et remplissent deux rôles différents lorsqu'ils cherchent à les surmonter. D'une part, les gouvernements sont appelés à réguler les algorithmes pour protéger la société des effets et préjudices potentiels que l'IA pourrait engendrer à court et long terme. Les responsables politiques jouent à ce titre un rôle clé dans la régulation et la gouvernance des usages de l'IA par les entreprises privées. D'autre part, les gouvernements veulent eux aussi profiter des potentialités de l'IA pour optimiser leur fonctionnement en intégrant l'IA au sein des ministères et des organismes publics (Paul et collab., 2024 ; Wirtz et Muller, 2019).

Au cours des dernières années, la plupart des experts dans le domaine de l'IA ont essentiellement positionné l'État en tant que *régulateur* ou, au mieux, en tant que *facilitateur* établissant les conditions optimales pour une utilisation responsable de l'IA par les entreprises privées et la population. Par conséquent, le rôle du secteur public en tant que grand *consommateur* et *bénéficiaire* de l'IA est bien souvent sous-estimé. Il est d'ailleurs intéressant de constater que l'orientation actuelle des débats entourant l'IA se concentre davantage sur la gouvernance *de* l'IA (les autorités publiques occupant un rôle de *régulateur* et *facilitateur* du développement et de l'utilisation de l'IA) et moins sur la gouvernance *avec* l'IA (les autorités publiques comme *consommatrices* et *bénéficiaires* de l'IA).

Ainsi, lorsque les responsables politiques ont décidé d'allouer des budgets considérables pour exploiter le potentiel de l'IA dans le développement économique, ils ont adopté des stratégies gouvernementales visant à planifier et à orienter le financement de la recherche et de l'innovation dans ce domaine considéré comme stratégique pour la croissance économique. Malgré le potentiel de l'IA pour accroître l'efficacité des politiques et améliorer la qualité des services publics, de nombreuses stratégies nationales en matière d'IA adoptées à l'échelle internationale ont souvent négligé ou abordé de manière superficielle le rôle de l'État en tant qu'utilisateur de cette technologie (Kuziemski et Misuraca, 2020 ; Jacob et Lawarée, 2022). Certaines initiatives récentes témoignent, tout de même, d'une volonté de reconnaître plus explicitement ce rôle. En effet, au début des années 2020, quelques pays se sont dotés de stratégies spécifiques relatives à l'intégration de l'IA dans le secteur public. C'est ainsi qu'en 2021 le gouvernement suédois a mandaté l'Agence pour l'administration numérique (Myndigheten för digital förvaltning) de coordonner les initiatives visant à accroître la capacité de l'administration publique à utiliser l'IA (OECD.AI, 2021). Au même moment, le gouvernement du Québec a adopté la Stratégie d'intégration de l'intelligence artificielle dans l'administration publique (Gouvernement du Québec, 2021).

À une époque où les questions de justice sociale et d'inclusion occupent une place centrale dans nos sociétés, il est essentiel de comprendre comment l'IA peut être utilisée par les organisations publiques pour concevoir et fournir des services publics justes, équitables et exempts de discrimination. Pour y parvenir, il est nécessaire de décoder les biais algorithmiques. C'est à cela que nous nous employons dans cet ouvrage.

1. DÉCRYPTER LE FONCTIONNEMENT DE L'INTELLIGENCE ARTIFICIELLE

L'intelligence artificielle[3] désigne « un dispositif technique traduisant son environnement sous forme de données afin d'inférer des structures et/ou de générer du contenu ayant un sens en fonction d'un objectif déterminé » (Bertolucci, 2024, p. 74). En combinant plusieurs applications d'IA au sein d'un système d'intelligence artificielle (SIA), les ordinateurs peuvent accomplir des tâches complexes, telles que la collecte et l'interprétation de données ou le raisonnement pour déterminer les meilleures actions à entreprendre (Wirtz et Muller, 2019 ; AI HLEG, 2018).

Par exemple, le système Automated Bus Lane Enforcement (ABLE) est un SIA qui, à New York, utilise des caméras installées sur les autobus pour surveiller la circulation dans les couloirs réservés afin de favoriser la fluidité de la circulation du transport en commun. Les caméras permettent tout d'abord de repérer les véhicules circulant dans les couloirs d'autobus grâce à la vision par ordinateur. En utilisant le GPS et des technologies de géorepérage, ABLE est alors capable de déterminer avec précision la position d'un autobus et de vérifier si le véhicule qui le précède est autorisé ou non à se trouver dans le couloir réservé (par exemple, pour tourner

3. Comme nous allons le voir dans cette section, les notions de base de l'IA comprennent des termes qui ont un sens particulier, tout en étant relativement interreliés. Pour faciliter la lecture de cet ouvrage, nous emploierons comme synonymes les termes « algorithme », « intelligence artificielle (IA) », « système d'aide de décision automatisée » et « système d'intelligence artificielle (SIA) ». Nous réserverons cependant l'emploi du terme IA aux cas les plus généraux et nous considérons qu'un algorithme est en soi un « système », constituant ou étant une partie constituante d'un système plus large (SIA ou système d'aide à la décision automatisée).

à droite à une intersection). Si le véhicule est en infraction, ABLE capte alors sa plaque d'immatriculation et les photographies de la plaque sont nettoyées avec des technologies de traitement des images et de reconnaissance optique de caractères. Finalement, les données collectées (images, vidéos, informations de localisation) sont transmises, en temps réel, au Département des transports de la ville de New York pour le traitement du dossier et l'envoi du constat d'infraction (Jacob et Brousseau, 2024).

Parmi les termes en vogue en matière d'intelligence artificielle, l'algorithme est le plus populaire. « Si vous regardez autour de vous, vous trouverez des algorithmes dans les endroits les plus ordinaires : votre four à micro-ondes, votre machine à laver et, bien sûr, votre ordinateur. Vous demandez à des algorithmes de vous faire des recommandations : quelle musique vous plairait ou quel itinéraire faut-il emprunter en voiture. […] Le mot "algorithme" figure dans l'actualité presque tous les jours » (Cormen et collab., 2022, p. XIII, trad.). Mais, concrètement, qu'est-ce qu'un algorithme ? La réponse à cette question se trouve dans l'encadré 1.

Il existe plusieurs types d'IA conçus pour accomplir des tâches spécifiques en fonction de leurs caractéristiques. C'est un peu comme lorsque nous ouvrons un logiciel pour écrire un texte sur notre ordinateur, un autre pour envoyer des courriels, gérer le budget familial ou éditer des photos. De la même manière, les applications d'IA possèdent leurs propres spécificités et peuvent être regroupées en grandes catégories. L'IA générative, par exemple, est devenue très populaire grâce à ChatGPT, qui permet de créer du texte, des images et même de la musique ou des vidéos. D'un autre côté, l'apprentissage automatique reste un pilier de l'IA

permettant aux systèmes informatiques de s'améliorer et de gagner en précision grâce à leurs facultés d'apprentissage à partir des données qu'ils analysent. Ce type d'IA est utilisé principalement pour des tâches telles que la reconnaissance d'images, la prédiction de tendances et l'automatisation de processus.

ENCADRÉ 1
Qu'est-ce qu'un algorithme ?

« Un algorithme est la description d'une suite d'étapes permettant d'obtenir un résultat à partir d'éléments fournis en entrée. Par exemple, une recette de cuisine est un algorithme permettant d'obtenir un plat à partir de ses ingrédients ! Dans le monde, de plus en plus numérique, dans lequel nous vivons, les algorithmes mathématiques permettent de combiner les informations les plus diverses pour produire une grande variété de résultats : simuler l'évolution de la propagation de la grippe en hiver, recommander des livres à des clients sur la base des choix déjà effectués par d'autres clients, comparer des images numériques de visages ou d'empreintes digitales, piloter de façon autonome des automobiles ou des sondes spatiales, etc.

Pour qu'un algorithme puisse être mis en œuvre par un ordinateur, il faut qu'il soit exprimé dans un langage informatique, sous la forme d'un logiciel (aussi appelé "application"). Un logiciel combine en général de nombreux algorithmes : pour la saisie des données, le calcul du résultat, leur affichage, la communication avec d'autres logiciels, etc.

Certains algorithmes ont été conçus de sorte que leur comportement évolue dans le temps, en fonction des données qui leur ont été fournies. Ces algorithmes "auto-apprenants" relèvent du domaine de recherche des systèmes experts et de l'"intelligence artificielle". Ils sont utilisés dans un nombre croissant de domaines, allant de la prédiction du trafic routier à l'analyse d'images médicales » (*Algorithme | CNIL*, s. d.).

L'apprentissage automatique repose sur la détection et l'exploitation des corrélations statistiques dans les données. Ce type d'IA, centré sur l'art de décider et de modéliser à partir de données, a émergé à la fin des années 1950. À ses débuts, l'apprentissage automatique se concentrait principalement sur des tâches statiques de classification, comme la reconnaissance de l'écriture manuscrite. Cependant, les techniques d'apprentissage automatique sont rapidement tombées en disgrâce et ont connu un regain d'intérêt au cours des dernières décennies, sous la forme de réseaux de neurones profonds (*deep neural networks*), grâce à l'expansion des capacités de calculs et à la disponibilité de grandes quantités de données. En effet, les algorithmes d'apprentissage automatique ont souvent besoin de grands ensembles de données pour fonctionner efficacement, ce qui est rendu possible avec l'émergence des données massives (*big data*) (Dean et collab., 2021). Ces données collectées à grande échelle proviennent de diverses sources, telles que les réseaux sociaux, les appareils connectés ou les téléphones cellulaires. Les grandes entreprises technologiques, comme Google, Apple, Facebook, Amazon et Microsoft[4], analysent en permanence les activités de leurs utilisateurs pour collecter des données et créer des profils détaillés, permettant de cibler les publicités que nous recevons, mais également d'orienter nos comportements et nos choix personnels (Zuboff, 2020).

Les algorithmes d'apprentissage automatique, notamment ceux qui utilisent les réseaux de neurones profonds, sont capables d'accomplir des prouesses remarquables. Nous sommes arrivés à une situation à la fois fascinante

4. L'acronyme GAFAM est souvent utilisé pour parler de ces cinq grandes compagnies technologiques.

et inquiétante où « les données massives et l'IA permettent aux ordinateurs de comprendre le monde comme nous, les humains, ne pouvons pas le faire » (Petrogradskaya et collab., 2021, p. 716, trad.). Ces algorithmes présentent néanmoins un défaut important : leur fonctionnement est opaque. Ce phénomène est décrit par les spécialistes comme le problème de la boîte noire (*black box*), c'est-à-dire que, dans certains cas, on n'est pas capable de comprendre comment l'IA est arrivée à produire une décision ou une prédiction, car il n'est pas possible de connaître les critères qui ont été pris en compte pour parvenir au résultat ou de retracer les étapes du processus de raisonnement. Ce manque d'explicabilité est particulièrement problématique quand l'IA est utilisée au sein d'organisations publiques et qu'elle alimente les processus de prise de décision. Pour répondre à ces problèmes, la communauté de recherche en IA s'est concentrée sur l'explicabilité des modèles afin de concevoir une « IA de confiance ». Nous en reparlerons au chapitre 3.

L'optimisation de la précision prédictive des systèmes d'IA n'est cependant pas systématiquement synonyme de résultats bénéfiques pour tout le monde. À cet égard, la sensibilisation aux biais potentiels des systèmes algorithmiques s'est accentuée ces dernières années, en particulier en ce qui a trait à l'exactitude et la représentativité des prédictions et décisions algorithmiques, à la suite de la médiatisation de problèmes majeurs de sexisme ou de racisme sur lesquels nous reviendrons. L'importance accordée à l'équité, en lien avec l'utilisation des algorithmes d'apprentissage automatique, a conduit à l'émergence de conférences spécialisées sur le sujet, telles que la Conférence annuelle sur l'équité, la responsabilité et la transparence (Conference on Fairness, Accountability, and Transparency), instaurée en 2018 par l'Association for Computing Machinery (ACM) (Dean et

collab., 2021). Cependant, même si les techniques mathématiques sont utiles pour formaliser les incompatibilités et les compromis algorithmiques, il devient de plus en plus évident que les choix concernant la manière de résoudre les biais algorithmiques nécessitent des raisonnements et des considérations philosophiques et sociales qui dépassent le cadre technique étroit de l'approche standard de l'apprentissage automatique équitable (*fairness in machine learning – fair ML*) (Fazelpour et Danks, 2021). En effet, « l'équité et la justice sont des propriétés des systèmes sociaux […] et non des propriétés des outils technologiques [utilisés au sein de ces systèmes sociaux] » (Selbst et collab., 2019, p. 59, trad.).

Les défis soulevés par l'IA et les décisions basées sur les algorithmes semblent récents et largement médiatisés. Cependant, ils s'inscrivent dans le prolongement du développement des systèmes traditionnels de technologies de l'information (TI). Les algorithmes d'apprentissage automatique se distinguent toutefois par leurs capacités à apprendre et à s'autonomiser, ce qui risque d'engendrer des erreurs susceptibles d'avoir des conséquences plus importantes sur la vie de certaines personnes ou communautés. Contrairement aux générations précédentes d'algorithmes, basés sur des règles, qui suivent des instructions spécifiques établies par des spécialistes et qui répètent leurs éventuelles erreurs de programmation, les algorithmes d'apprentissage automatique plus récents peuvent quant à eux commettre de nouvelles erreurs qui sont difficiles, voire impossibles, à détecter. Une grande partie des nouveaux défis associés à l'IA découle des caractéristiques propres aux algorithmes d'apprentissage automatique, qui les différencient des solutions numériques utilisées depuis longtemps par les organisations publiques. Ces caractéristiques incluent une puissance de calcul surpassant les capacités cognitives

humaines, l'apprentissage autonome, c'est-à-dire la création de connaissances sans supervision directe, le profilage par catégorisation des traits et des comportements ainsi que l'incitation comportementale (*nudging*) qui contribue au respect des règles ou oriente certains comportements. Ces éléments génèrent des enjeux absents de l'IA basée sur des règles des époques antérieures (Kuziemski et Misuraca, 2020). Dans le secteur public, la majorité des biais algorithmiques répertoriés surviennent dans des situations où de vastes ensembles de données historiques reflétant le passé ont été utilisés pour entraîner des modèles prédictifs au moyen de méthodes algorithmiques ou statistiques. Dans ce contexte, le terme « algorithmique » désigne généralement les systèmes automatisés de prise de décision.

L'administration publique remplit plusieurs fonctions, telles que la gestion des ressources, la mise en œuvre des politiques publiques et la fourniture de services à la population. Parmi ces responsabilités, la prise de décision occupe une place centrale. Il est donc compréhensible que l'automatisation de la prise de décision soit envisagée par les organisations publiques soucieuses d'améliorer leurs performances à moindre coûts.

2. PRÉVENIR LES BIAIS ALGORITHMIQUES : IL EST TEMPS D'AGIR !

Comme la volonté de réduire les coûts et d'optimiser l'efficacité des services publics entraîne une adoption rapide des technologies d'IA dans les organisations publiques, il est essentiel de prévoir les défis et les problèmes potentiels que l'IA pourrait générer avant qu'ils surviennent, afin d'éviter de devoir réagir en urgence ou d'en subir les effets néfastes. Il faut reconnaître que les avancées des technologies d'IA

progressent à un rythme effréné, laissant peu de temps pour entreprendre une réflexion approfondie nécessaire à l'établissement d'une réglementation appropriée. Ce décalage entre réflexion et innovation est manifeste en ce qui concerne la question des biais algorithmiques. L'accélération du développement de l'IA pourrait nous amener à un moment où la réduction ou l'élimination des répercussions négatives de cette technologie deviendront plus difficiles et peut-être même impossibles. Pour l'instant, nous avons encore la possibilité de contrôler et de choisir les utilisations de l'IA qui correspondent aux valeurs que nous souhaitons promouvoir et protéger (Bannister et Connolly, 2020).

Les problèmes associés aux algorithmes de prise de décision et d'assistance à la décision automatisés sont, en grande partie, techniques. Ils découlent souvent de la conception, du développement et du déploiement de l'IA. Le rôle des concepteurs et des évaluateurs de ces systèmes est donc indispensable afin de comprendre les causes, de repérer les sources et de résoudre les problèmes techniques à l'origine des biais algorithmiques. Il serait cependant erroné de penser que le problème est purement technique et qu'il sera complètement résolu grâce à l'optimisation des modèles ou aux progrès technologiques. Une gouvernance efficace et une régulation appropriée sont donc nécessaires. Pour cela, il est essentiel de comprendre clairement les aspects à réglementer et de favoriser une collaboration interdisciplinaire entre les chercheurs en sciences naturelles, en sciences informatiques, en génie et en sciences humaines et sociales.

En effet les difficultés relatives à l'intégration des travaux existants sur les biais algorithmiques ne découlent pas uniquement de la rapidité du développement technologique, mais aussi d'une définition imprécise du problème. Le terme « biais algorithmique », bien qu'il fasse référence à un

problème principalement technique, englobe une variété de dimensions – statistique, historique, culturelle ou linguistique – dans un contexte où la littératie numérique est souvent insuffisante et où les questions de justice, d'équité et de discrimination sont fortement médiatisées (Selbst et collab., 2019)[5].

La question des biais algorithmiques est complexe et multidimensionnelle. L'objectif de cet ouvrage est d'offrir une synthèse des connaissances sur les risques de biais algorithmiques dans le secteur public et de présenter des solutions en vue de les prévenir, de les atténuer, de les corriger ou de les supprimer. Il est en effet important de distinguer aussi précisément que possible les enjeux techniques et socioculturels associés à l'IA afin d'assurer une meilleure compréhension de son fonctionnement et de ses répercussions, et de favoriser une prise en charge adéquate des risques comme des occasions de l'intégration de cette technologie dans le secteur public. Les enjeux et les risques que nous présentons dans les pages suivantes concernent autant les organisations publiques que privées. Toutes ces organisations sont encouragées à mettre en place des balises

5. Une autre approche, plus critique de l'IA en général, voit cette insistance sur le problème des biais algorithmiques comme l'occultation d'un problème plus large. Selon cette perspective critique, « il existe deux histoires liées au développement de la technologie. L'une concerne l'intelligence artificielle, la promesse dorée [qui] est présentée comme une force puissante, omniprésente et inarrêtable capable de résoudre nos plus grands problèmes […]. La seconde histoire est que l'IA a un problème : les biais. Les exemples de partialité sont légion : des publicités en ligne qui proposent aux hommes des emplois mieux rémunérés ; des services de livraison qui évitent les quartiers défavorisés ; des systèmes de reconnaissance faciale qui ne prennent pas en compte les personnes de couleur ; des outils de recrutement qui éliminent les femmes de manière invisible. […] En acceptant ces récits sur l'IA, […] ce que l'on obtient, c'est la résignation – la normalisation de la capture massive de données, un transfert à sens unique vers les entreprises technologiques et l'application de solutions automatisées et prédictives à tous les problèmes de la société » (Powles et Nissenbaum, 2018, p. 2-3, trad.).

et des mécanismes visant à réduire ou à supprimer les biais algorithmiques. Notre attention se concentrera cependant davantage sur les organisations publiques[6] en raison du rôle particulier qu'elles remplissent dans la production du bien commun et la fourniture de services à la population. En effet, la prise de décision automatisée peut engendrer des répercussions majeures sur la population ou des groupes spécifiques de la société.

Cet ouvrage est divisé en trois chapitres. Le premier chapitre décrit le contexte de l'utilisation des algorithmes d'apprentissage automatique par l'État. Il répond à deux questions :

- Pourquoi les organisations publiques utilisent-elles l'IA ?
- Comment les organisations publiques utilisent-elles l'IA ?

Nous commencerons par discuter des principales raisons qui expliquent le déploiement et l'utilisation de systèmes d'IA dans le secteur public. Nous illustrerons cette partie avec quelques exemples d'utilisations et d'applications de systèmes algorithmiques dans divers domaines d'action publique. Nous présenterons ensuite les préoccupations soulevées par l'utilisation de ces systèmes dans le secteur public, notamment en ce qui concerne ce que plusieurs qualifient de « gouvernance algorithmique ». Nous pourrons

6. La distinction entre les organisations publiques et privées est par ailleurs moins évidente à faire de nos jours en raison de l'évolution du fonctionnement de l'État depuis la mise en œuvre des réformes qui s'inspirent de la nouvelle gestion publique (NGP). Ainsi, certaines organisations qui fournissent des services publics disposent de statuts particuliers, comme les agences publiques et les sociétés d'État. Par ailleurs, en raison de la sous-traitance à des entreprises privées et des partenariats public-privé (PPP), certaines entreprises privées se voient confier des missions d'intérêt général.

constater que plusieurs de ces préoccupations découlent du problème des biais algorithmiques ou les exacerbent.

Dans le deuxième chapitre, nous examinerons plus en détail la question des biais algorithmiques en nous concentrant sur les raisons qui peuvent expliquer leur apparition et sur les moments où ils sont les plus propices d'apparaître dans le processus de développement et d'intégration de l'IA aussi appelé « cycle de vie » ou « pipeline de l'IA[7] ». En effet, même si les biais algorithmiques sont étroitement liés aux artefacts techniques[8], de multiples décisions individuelles et organisationnelles sont prises lors de la spécification du problème à résoudre, de la conception et du déploiement des algorithmes d'apprentissage automatique. Ces décisions reflètent les préférences, les intérêts et les valeurs individuelles, organisationnelles ou sociétales et vont ultimement teinter l'utilisation, le fonctionnement, la performance et les résultats de l'algorithme développé. Dans le deuxième chapitre, nous examinerons aussi les répercussions des biais algorithmiques sur le secteur public, ainsi que plus largement les risques d'injustice et de discrimination résultant de l'utilisation de ces outils dans la société.

Dans le troisième chapitre, nous décrirons les pistes de solution permettant aux organisations publiques de prévenir, réduire ou supprimer les biais algorithmiques. Nous présenterons des actions pratiques et des recommandations concrètes pour répondre aux problèmes décrits dans les

7. « Le pipeline inclut la collecte des données, l'intégration de celles-ci dans des fichiers de données d'apprentissage, l'entraînement d'un ou plusieurs modèles et l'exportation des modèles en production » (DataFranca, 2024).
8. En apprentissage automatique, le terme « artefact » désigne tous les produits numériques utilisés ou produits pour développer un modèle algorithmique, tels que les données d'entraînement, le fichier créé au cours du processus d'entraînement, le modèle final, etc.

chapitres précédents. Nous constaterons que, dans certains cas, il est impossible d'éliminer complètement les biais. Une approche suggérée vise alors à réduire au minimum l'amplification des biais par les systèmes d'IA (OECD, 2024). Cela requiert, par exemple, de prêter attention à la raison d'être, à la finalité et au contexte dans lequel un système algorithmique est conçu et utilisé.

CHAPITRE 1

L'intelligence artificielle dans le secteur public

L'intégration de l'IA dans le secteur public est la manifestation la plus récente de la transformation numérique de l'État entamée par de nombreux gouvernements soucieux de se moderniser, en tirant avantage des progrès fulgurants des nouvelles technologies. La montée en puissance de l'IA dans le secteur public est souvent qualifiée de révolutionnaire et s'apparente à une baguette magique qui serait capable de corriger les maux traditionnels de la bureaucratie, tels que la lourdeur, la lenteur et la rigidité (Bullock, 2019 ; Busch et Henriksen, 2018 ; Paul et collab., 2024 ; Young et collab., 2019 ; Wirtz et Muller, 2019). Ces ambitions se retrouvent d'ailleurs dans les explications fournies par le ministère de la Cybersécurité et du Numérique (MCN) du gouvernement du Québec lorsqu'il présente sa stratégie d'intégration de l'IA dans l'administration publique. Selon le MCN,

> le recours à l'IA par le secteur public peut notamment contribuer à :
>
> - Prodiguer des conseils et des services qui correspondent mieux aux besoins des citoyens dans différentes situations de vie ;
> - Offrir un meilleur soutien à la prise de décision de ses employés ;
> - Rationaliser les processus et optimiser l'utilisation des ressources, notamment par l'automatisation de certaines tâches répétitives ;

- Améliorer la qualité des processus et des services en détectant automatiquement les anomalies ;
- Établir des tendances et avancer des prédictions en s'appuyant sur de grandes quantités de données (MCN, 2024a, p. 4).

Les systèmes d'intelligence artificielle sont donc promus et adoptés pour leur efficacité, leur rapidité et leur neutralité. Des chercheurs expliquent que les systèmes de prise de décision basés sur les algorithmes sont perçus comme supérieurs aux processus décisionnels impliquant des humains, car les algorithmes sont immunisés contre les biais psychologiques et qu'ils parviennent ainsi à prendre des décisions adéquates et impartiales (König et Wenzelburger, 2021). À une époque où les préoccupations relatives à la discrimination et aux principes d'équité, de diversité et d'inclusion (EDI) occupent une place importante dans le débat public, l'utilisation de l'IA est considérée, par certains chercheurs, comme une solution permettant d'objectiver les décisions prises par les organisations publiques (Ferguson, 2017).

Lorsqu'on s'intéresse aux biais algorithmiques dans le secteur public, nous avons tendance à considérer que l'IA est la seule à être à l'origine des problèmes d'injustice et de discrimination (Miller, 2018). Nous avons tendance à oublier que les fonctionnaires ont eux aussi des préjugés à l'égard des usagers ou que les organisations publiques sont à l'origine de pratiques discriminatoires ou d'injustices systémiques à l'égard de certains groupes de la population. Les agents de première ligne (*street-level bureaucrats*), tels que les travailleurs sociaux, les policiers ou les fonctionnaires au guichet, disposent d'une marge de manœuvre aussi qualifiée

de « pouvoir discrétionnaire[1] ». C'est grâce à ce pouvoir que les agents publics sont en mesure de prendre des décisions adaptées aux caractéristiques spécifiques des personnes et du contexte dans lequel elles se trouvent. Bien que les agents de première ligne s'efforcent de traiter des cas similaires de la même manière, leurs décisions doivent souvent s'adapter à des circonstances spécifiques, ce qui risque d'introduire des incohérences et des biais (Scott, 1997 ; Raaphorst, 2021). Deux fonctionnaires pourraient ainsi prendre des décisions différentes sur le même dossier en fonction de leur interprétation des politiques, de leurs préjugés, des caractéristiques des usagers ou des formes de contrôle organisationnel (Hupe et Hill, 2007). Tous ces éléments peuvent nuire à l'équité, à la cohérence des décisions et même aboutir à des biais.

Avec la transformation numérique de l'administration, de plus en plus d'interactions administratives se font à l'aide d'outils technologiques. Le déploiement de ces outils suscite de nombreuses interrogations et critiques, car ils risquent de déshumaniser la fourniture de service ou d'affaiblir le jugement professionnel des agents publics. Ainsi, les travaux sur la transformation numérique de l'administration publique ont fait apparaître de nouvelles expressions, telles que « désert numérique » pour décrire une zone où l'accès à Internet est insuffisant ou inexistant, « fracture numérique » ou « orphelin numérique » pour désigner les personnes ou les groupes marginalisés par leur manque d'accès aux

1. Dans la littérature en administration publique, le terme « pouvoir discrétionnaire » désigne la capacité d'une personne à prendre des décisions en ayant une marge de manœuvre pour apprécier la situation. Ce type de pouvoir permet de choisir entre différentes options selon le contexte et les besoins spécifiques. Par exemple, un juge peut adapter la sanction en fonction des circonstances particulières d'un cas, tant qu'il reste dans les limites fixées par la loi. Ce pouvoir repose donc sur le jugement personnel de l'agent public.

technologies numériques ou par une maîtrise insuffisante des outils numériques. D'autres études nous apprennent que l'automatisation et les outils technologiques offrent des solutions à des problèmes, tels que les mauvaises décisions dues à des erreurs ou à une interprétation erronée des règles, le favoritisme, la discrimination ou la corruption (Busch et Henriksen, 2018).

L'IA peut aussi être mise au service de la quête de fiabilité et d'objectivité des organisations publiques (Kuziemski et Misuraca, 2020). Les décisions et les prédictions algorithmiques qui se fondent sur des données et des analyses statistiques sont bien souvent perçues comme objectives et neutres. Des études suggèrent que l'adoption des systèmes algorithmiques dans différents domaines d'action publique est motivée par ces promesses de décisions plus objectives et exemptes des préjugés humains (Ferguson, 2017). Dans cette optique, ces systèmes peuvent devenir des instruments d'aide à la prise de décision ou être capables de déceler des biais dans les décisions prises par les organisations ou les agents publics (Cluzel-Métayer, 2020).

1. POURQUOI LES ORGANISATIONS PUBLIQUES UTILISENT-ELLES L'IA ?

Les systèmes de décision automatisée dans le secteur public, combinés aux capacités grandissantes de transfert et de traitement de données, offrent des solutions pour accomplir des tâches complexes à une échelle inédite. La quête d'innovation est un incitatif majeur qui pousse les gouvernements à intégrer des solutions technologiques avancées. Dans certains cas, c'est d'ailleurs principalement la recherche de la nouveauté et l'attrait de la modernité qui poussent les gouvernements à adopter des solutions

technologiques. Il y a quelques années, le dirigeant d'un organisme public nous indiquait, lors d'une entrevue réalisée dans le cadre d'un projet de recherche, que « l'IA s'apparente à une solution à la recherche d'un problème ». Étant donné que de nombreux gouvernements sur la planète cherchent à tirer profit du potentiel de l'IA, certains responsables politiques vont alors avoir tendance à formuler des promesses excessives sur les avantages de la technologie pour en faciliter l'adoption ou vouloir se hisser, coûte que coûte, au sommet des classements internationaux qui mesurent l'adoption ou la maturité numérique des pays. De telles dynamiques engendrent des discours compétitifs, similaires à ceux d'une « course aux armements », qui se reflètent dans les choix politiques des décideurs.

Les algorithmes, en analysant d'immenses quantités de données, révèlent des corrélations souvent insaisissables par les humains. Des organisations, telles que la Banque mondiale ou le Government Accountability Office (GAO) aux États-Unis, ont récemment mis au point des solutions d'IA pour explorer les vastes quantités de rapports qu'elles ont publiés au fil du temps. Le but de l'utilisation de l'IA dans ce contexte est de synthétiser les connaissances internes déjà produites pour générer de nouvelles connaissances (Jacob, 2024). La Banque mondiale a utilisé de l'apprentissage machine non supervisé[2] pour analyser près de 400 rapports concernant 64 pays bénéficiaires de l'aide internationale,

2. « En apprentissage non supervisé, l'algorithme d'apprentissage automatique découvre des régularités statistiques, des formes ou des structures dans des données qui ne comportent pas d'annotation (ou étiquette). Pour y arriver, l'apprentissage non supervisé se fonde sur la détection de similarités entre les données. Dans cette approche, le nombre de classes [c'est-à-dire de catégories] et leur nature ne sont pas nécessairement prédéterminés, c'est l'algorithme qui les découvrira en fonction des données analysées » (DataFranca, 2024).

en vue de déterminer les facteurs de succès et d'échec des projets. À la fin de cette expérimentation, elle a estimé que l'approche a permis de connaître les facteurs qui « étaient non seulement cohérents et pertinents pour le domaine, mais aussi originaux et judicieux. Cela a ajouté une profondeur supplémentaire à l'analyse en distinguant des facteurs qu'une analyse réalisée par un humain aurait pu manquer » (Franzen et collab., 2022, p. 27, trad.). Avec cet exemple, nous voyons que l'IA dans le secteur public permet de rassembler ou de développer des connaissances susceptibles d'alimenter les réflexions lors de l'élaboration, de la mise en œuvre ou de l'évaluation des politiques publiques (Jacob, 2025).

Alors qu'en principe la prise de décision algorithmique peut être entièrement automatisée, les algorithmes d'IA sont utilisés surtout pour assister l'humain dans la prise de décision, plutôt que de le supplanter complètement. Cela est particulièrement vrai dans le secteur public, où l'automatisation complète semble inappropriée ou inatteignable, en particulier dans les cas de systèmes d'IA qui peuvent potentiellement affecter la vie des gens de manière significative (Engstrom et collab., 2020 ; Jacob et Brousseau, 2024). Ainsi, bien que la neutralité perçue de la prise de décision algorithmique, comparée à la subjectivité ou à l'intuition humaine, ait été un catalyseur majeur de son intégration dans le secteur public, les systèmes algorithmiques jouent davantage, à l'heure actuelle, le rôle de soutien décisionnel que celui de décideurs autonomes à part entière, en raison notamment de contraintes juridiques (Alon-Barkat et Busuioc, 2023). Les systèmes algorithmiques sont donc utilisés surtout dans un contexte où « une sortie [*output*], une prédiction ou une recommandation générée par un algorithme au moyen d'une technique d'apprentissage automatique est intégrée dans un

processus décisionnel nécessitant l'approbation ou la mise en œuvre par un humain » (Oswald, 2018, p. 2-3, trad.).

Cependant, que la décision soit prise uniquement par un système d'IA ou qu'elle guide celle d'un agent public, la question de la responsabilité demeure la même. Qui doit être tenu responsable en cas d'erreur administrative résultant de l'utilisation de l'IA dans le processus administratif ? S'agit-il des concepteurs, des programmeurs, des gestionnaires ou des agents de première ligne ? À l'heure actuelle, ces questions sont sans réponse claire. Certaines juridictions, comme l'Union européenne, au moyen du *Règlement général sur la protection des données* (RGPD), encadrent cet usage en « mettant de l'avant le droit […] de ne pas être soumis à une décision basée uniquement sur un processus décisionnel automatisé » (Busuioc, 2021, p. 828, trad.). Au Québec, ce principe est également promu dans les principes de mise en œuvre de la Stratégie d'intégration de l'intelligence artificielle dans l'administration publique, qui stipule que « la prise de décision demeure sous la responsabilité du personnel de l'État » (MCN, 2024b). Cela signifie qu'une décision ne peut être prise uniquement par une IA sans supervision ou validation humaine.

2. COMMENT LES ORGANISATIONS PUBLIQUES UTILISENT-ELLES L'IA ?

Les algorithmes d'intelligence artificielle jouent un rôle prépondérant dans de très nombreux domaines d'action publique (Diakopoulos, 2015 ; Eubanks, 2018 ; Paul et collab., 2024 ; Yeung, 2018 ; Veale et Brass, 2019 ; Busuioc, 2021). Ils sont notamment utilisés pour générer des prédictions et alimenter des processus décisionnels complexes qui ont souvent des conséquences importantes pour les

personnes concernées. Par exemple, dans le système éducatif, des algorithmes prédisent les performances scolaires des étudiantes et des étudiants dans le processus d'admission à l'université. D'autres systèmes peuvent évaluer la performance des enseignants et formuler des recommandations quant à leurs compétences, en allant jusqu'à suggérer le licenciement de certains professeurs dont les performances sont jugées insatisfaisantes. Il y a quelques années, le district scolaire de Houston au Texas a été poursuivi en justice par plusieurs professeurs qui contestaient leur licenciement en invoquant le fait que le système d'IA utilisé pour calculer leur contribution à la progression scolaire des élèves était opaque et ne leur permettait pas de comprendre comment leurs notes étaient calculées ni d'en vérifier l'exactitude. Le district scolaire a conclu un accord à l'amiable avec les plaignants et a renoncé à utiliser cet algorithme en 2017, après qu'un juge fédéral eut permis la poursuite du procès (Paige et Amrein-Beardsley, 2020). Dans le secteur de la justice, des algorithmes influencent les paramètres de la remise en liberté et peuvent aussi être employés pour déterminer la nature des sanctions imposées à une personne reconnue coupable d'un crime ou d'un délit (Angwin et collab., 2016). Pour la police, les algorithmes de reconnaissance faciale peuvent faciliter les enquêtes, tandis que des algorithmes d'apprentissage automatique contribuent à prédire la probabilité que des délits ou des activités criminelles se produisent dans certains quartiers (Ferguson, 2017).

Dans le secteur public, l'IA remplit diverses fonctions qui profitent à différents bénéficiaires. Pour les usagers qui interagissent avec les organisations publiques ou qui utilisent les services gouvernementaux, l'IA permet d'accélérer les processus administratifs ou de rendre les services plus accessibles ou personnalisés. Pour les employés du secteur

public, l'IA permet d'automatiser des tâches répétitives et fastidieuses ou de les assister dans la prise de décision en analysant des documents ou formulant des recommandations. Les organisations publiques peuvent également bénéficier des systèmes d'IA dans leur fonctionnement global afin d'optimiser leurs processus internes et de les orienter dans leurs choix stratégiques (Desouza et collab. ; 2020, Jacob et Brousseau, 2024).

Afin d'illustrer différentes applications d'IA qui se déploient dans de nombreuses administrations, nous présentons dans l'encadré 2 quelques cas d'utilisation de l'IA dans le secteur public.

ENCADRÉ 2
Cas d'utilisation de l'IA dans le secteur public

Éducation – L'intégration des algorithmes dans les établissements d'enseignement transforme de nombreuses tâches. L'IA permet, entre autres, de prédire le risque de décrochage scolaire, de corriger les devoirs des élèves, d'évaluer les acquis des apprenants et de gérer les procédures d'admission aux études supérieures (Baker et Hawn, 2022 ; Kemper et collab., 2020 ; Ramineni et Williamson, 2013 ; Ritter et collab., 2016).

Emploi – Les organisations qui aident les personnes à trouver un emploi tirent elles aussi profit des technologies d'IA. En France, le système d'IA La Bonne Boîte aide les personnes en recherche d'emploi à trouver des employeurs potentiels. Ce système d'IA analyse des données provenant d'offres d'emploi, des tendances du marché du travail et des informations que les entreprises doivent transmettre à l'État. À l'aide d'algorithmes prédictifs, le système d'IA prévoit quelles entreprises sont susceptibles de recruter, même en l'absence d'offres d'emploi, ce qui permet aux personnes en recherche d'emploi d'envoyer des candidatures spontanées. De leur côté, les personnes en recherche d'emploi ont préalablement saisi dans la plateforme leurs préférences, leurs compétences et leur lieu de résidence. C'est en combinant toutes ces informations que le système

propose des recommandations personnalisées d'entreprises correspondant au profil des personnes en recherche d'emploi (Jacob et Brousseau, 2024).

Fiscalité – Les services des impôts en France utilisent depuis quelques années des outils technologiques combinés à l'IA pour détecter les piscines non déclarées en analysant des images aériennes. Un algorithme compare ces images aux bases de données fiscales pour repérer des piscines qui n'ont pas été enregistrées, donc qui ne sont pas taxées. Ce contrôle fiscal, augmenté par l'IA, permet de repérer rapidement des installations non déclarées et d'augmenter ainsi les recettes fiscales (DEWG, 2024). En 2024, le ministère des Comptes publics indiquait que l'utilisation de l'IA pour détecter les piscines non déclarées avait permis de repérer 140 000 installations non imposées. Grâce à cette opération, les collectivités locales ont pu récupérer 40 millions d'euros de recettes supplémentaires issues de la taxe foncière, sans compter les amendes que les propriétaires concernés risquent également de devoir payer pour avoir omis de déclarer ces équipements.

Immigration – L'IA peut améliorer les processus d'immigration. Des outils, tels que la reconnaissance faciale, l'analyse prédictive et la prise de décision automatisée, permettent d'optimiser le tri des demandes de visas, de faciliter les contrôles d'identité et de renforcer la sécurité aux frontières. Immigration Canada utilise des technologies d'IA pour améliorer les interactions des utilisateurs. Des robots conversationnels (*chatbots*) fournissent automatiquement des réponses aux demandes de renseignements. L'IA sert également à accélérer l'examen des dossiers, en vérifiant la conformité des informations contenues dans les formulaires électroniques avec les exigences institutionnelles (Kuziemski et Misuraca, 2020).

Justice – L'IA est également employée par les cours et les tribunaux d'un nombre croissant de pays. Les juges et les avocats peuvent utiliser l'IA pour faciliter la préparation d'un procès et pour entreprendre des recherches dans la jurisprudence. D'autres modèles prédictifs permettent aux avocats de la défense et aux procureurs d'estimer la probabilité de succès ou d'échec d'une affaire afin qu'ils puissent adapter leurs stratégies et leurs plaidoiries en conséquence. En automatisant la gestion du suivi des procédures, l'IA peut réduire l'arriéré judiciaire et améliorer

l'efficience des processus. Aux États-Unis, les juges disposent d'outils d'évaluation automatisée qui leur permettent de décider si un prévenu doit rester ou non en détention provisoire en attendant son procès. Des systèmes algorithmiques qui produisent des scores d'évaluation du risque de récidive sont aussi utilisés dans le système de justice pénale et aident les juges ou les commissions de libération à prendre une décision relative aux demandes de libération conditionnelle (Alon-Barkat et Busuioc, 2023).

Sécurité – Les forces de l'ordre adoptent des outils utilisant l'apprentissage automatique afin de les assister dans diverses tâches, notamment la reconnaissance faciale, l'analyse de vidéos, l'extraction de données mobiles, l'analyse de réseaux sociaux, la cartographie prédictive de la criminalité et l'évaluation des risques (Babuta et Oswald, 2019). Aux États-Unis, des systèmes de police prédictive analysent des données des crimes passés pour repérer des zones où pourraient survenir de nouveaux crimes ou délits. Les résultats des analyses prennent la forme de cartes thermiques, qui permettent aux responsables des services de police de visualiser les zones prioritaires lors de la répartition des patrouilles (Udoh, 2020). En 2024, l'arrestation de Daniella Klette, membre présumée de l'organisation allemande d'extrême gauche Fraction armée rouge (FAR), a fait les manchettes dans le monde entier. Des décennies de cavale ont pris fin grâce au travail d'un spécialiste en vérification d'image qui a utilisé la photo d'un vieil avis de recherche et un logiciel de reconnaissance faciale pour trouver sur Internet des photos correspondant à la fugitive la plus recherchée d'Allemagne.

Santé – Plusieurs applications d'intelligence artificielle transforment le secteur de la santé et le fonctionnement des hôpitaux. Ainsi, l'IA permet de diagnostiquer plus rapidement et plus efficacement certaines maladies en analysant les images médicales. Des systèmes d'IA aident les médecins à préparer des plans de traitement personnalisés qui s'appuient sur les dernières avancées de la science. Pour la gestion des hôpitaux, l'IA optimise la programmation et l'affectation des ressources en prédisant les taux d'admission des patients ou en surveillant l'état des stocks des fournitures médicales. De plus, des robots conversationnels et des assistants virtuels alimentés par l'IA permettent aux patients d'obtenir une assistance 24 heures sur 24 et 7 jours sur 7 (Henderson et collab., 2022).

Un élément commun à tous les exemples que nous venons de présenter est le fait que le système algorithmique, pour être efficace, va chercher à faire ressortir des caractéristiques ou des variables[3] dans les données afin d'établir des corrélations entre ces caractéristiques et les résultats ou les comportements attendus. Cela signifie qu'un système d'IA fonctionne en distinguant différents profils. Prenons le cas d'un système d'IA utilisé pour prédire la probabilité de retard ou de non-paiement d'impôts. En analysant des données telles que l'historique des paiements, le revenu, les déclarations précédentes et d'autres indicateurs financiers, l'IA peut identifier des contribuables présentant un risque plus élevé de ne pas respecter les délais de paiement. Ces personnes pourraient alors recevoir des rappels supplémentaires, des offres d'échelonnement de paiement ou même une assistance personnalisée pour éviter un non-paiement. Dans ce contexte, la discrimination faite par l'IA est utile parce qu'elle permet d'allouer les ressources publiques de manière plus ciblée. Plutôt que de traiter tous les contribuables de la même manière, l'administration peut concentrer ses efforts sur ceux qui ont besoin de soutien, tout en réduisant les coûts liés au recouvrement auprès de ceux qui paient régulièrement et à temps. Ainsi, comme le suggère un expert en matière de biais algorithmiques,

> [n]ous nous attendons à ce que l'algorithme soit biaisé par rapport aux personnes [en retard ou] qui ne peuvent pas payer, ce que la plupart des gens ne considèrent pas comme un problème. Mais,

3. Le terme « variable » fait référence à un élément ou une caractéristique mesurable qui peut varier d'un individu ou d'un cas à l'autre. Dans un modèle visant à prévoir le risque de maladies cardiaques, des variables comme l'âge, le poids et la pression artérielle pourraient être utilisées. Un algorithme de reconnaissance faciale utilise, quant à lui, les traits d'un visage, comme la distance entre les yeux ou la forme du nez pour différencier les visages et identifier les individus. En résumé, les variables aident l'algorithme à établir des relations et à faire des prédictions.

> si l'algorithme était biaisé en faveur des hommes ou des femmes, il s'agirait d'un biais [inacceptable]. Se débarrasser complètement des biais signifierait prendre des décisions au hasard, comme si l'on jouait à pile ou face, ce qui n'est pas la façon dont nous voulons que ces systèmes [algorithmiques] fonctionnent (IEEE Standards Association, 2023, trad.).

Cette précision est importante pour nous aider à comprendre la complexité du sujet, que nous allons tenter de décrypter dans les pages suivantes, en commençant par présenter les préoccupations liées à l'utilisation des systèmes algorithmiques dans le secteur public.

3. PRÉOCCUPATIONS SOULEVÉES PAR L'UTILISATION DES SYSTÈMES ALGORITHMIQUES DANS LE SECTEUR PUBLIC

Des débats et des polémiques émergent au fur et à mesure que les organisations publiques intègrent l'IA dans leur fonctionnement. L'encadré 3 recense six controverses emblématiques relatives à l'utilisation d'algorithmes d'aide à la décision dans le secteur public.

ENCADRÉ 3
Controverses entourant l'utilisation d'algorithmes dans le secteur public

2011 – Le service de police de Los Angeles a entrepris l'opération LASER (acronyme de Los Angeles Strategic Extraction and Restoration) avec l'objectif de cibler les délinquants multirécidivistes et de recenser les quartiers chauds de la ville pour déployer les effectifs policiers. Ce système de police prédictive a été abandonné en 2021, après avoir été critiqué pour avoir renforcé les préjugés raciaux. Le système ciblait de manière disproportionnée les communautés minoritaires, entraînant une présence policière et une surveillance accrue dans ces zones, sans que son efficacité ait été établie en matière de réduction de la criminalité.

2012 – Le ministère néo-zélandais du Développement social a été tenté de recourir à la modélisation prédictive pour identifier les enfants à risque de maltraitance et de négligence. Cette initiative visait à exploiter des données provenant de diverses sources, notamment des services sociaux, de la santé et de l'éducation, afin de prédire quels enfants étaient le plus à risque et pouvaient donc bénéficier d'une intervention précoce. Ses détracteurs ont fait valoir que le modèle risquait de porter atteinte à la vie privée et de stigmatiser les familles, en particulier celles qui sont issues de milieux minoritaires ou à faibles revenus. Anne Tolley, ministre du Développement social, a écrit ceci dans la marge d'un document décrivant la proposition : « pas sous ma gouverne, il s'agit d'enfants, pas de rats de laboratoire » (Jones, 2015, trad.).

2016 – Une enquête du journal *ProPublica* a dévoilé les biais raciaux produits par l'algorithme COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) utilisé par les services correctionnels américains pour estimer le risque de récidive de détenus sollicitant une libération. Les détracteurs de cet algorithme affirment que, sous le couvert d'une analyse objective des données, il renforce le racisme systémique et engendre des pratiques discriminatoires. En effet, le système COMPAS était biaisé à l'encontre des accusés afro-américains et prédisait des risques de récidive plus élevés pour eux que pour les accusés blancs (Angwin et collab., 2016).

2019 – Un juge australien déclare l'illégalité du système Robodebt, conçu pour évaluer et recouvrer automatiquement les trop-perçus et les dettes des bénéficiaires de l'aide sociale. Ce système, mis en place en 2016, a engendré de nombreuses erreurs et des personnes ont été accusées à tort de devoir de l'argent à l'État. Certaines de ces personnes ont dû contracter des emprunts, puiser dans leurs économies ou vendre leur voiture pour s'acquitter de la prétendue dette. Plusieurs personnes se sont même suicidées en raison de la détresse émotionnelle et financière qu'elles ont ressentie. La commissaire Holmes, qui a présidé une commission royale d'enquête sur ce scandale, a indiqué que « Robodebt était un mécanisme insensible et cruel, ni juste ni légal, qui donnait à de nombreuses personnes l'impression d'être des criminels » (Mao, 2023, trad.).

2020 – Pendant la pandémie de COVID-19, des centaines d'élèves anglais se sont rassemblés devant le ministère de l'Éducation en scandant : « *F**k the algorithm*![4] » La cible de leur colère était un algorithme que le gouvernement avait utilisé pour estimer les résultats du baccalauréat après l'annulation des examens. « Près de 40 % des étudiants ont obtenu des notes inférieures à celles qu'ils anticipaient, ce qui a déclenché un tollé et une action en justice » (Kolkman, 2020, trad.). L'algorithme a par ailleurs sous-évalué les notes des élèves issus de milieux défavorisés et augmenté celles des élèves inscrits dans des écoles privées. Pour calmer les esprits, le gouvernement britannique a finalement annoncé que les notes des élèves seraient basées sur l'estimation de leurs professeurs afin de refléter ce qu'auraient été leurs résultats si les examens avaient eu lieu comme prévu.

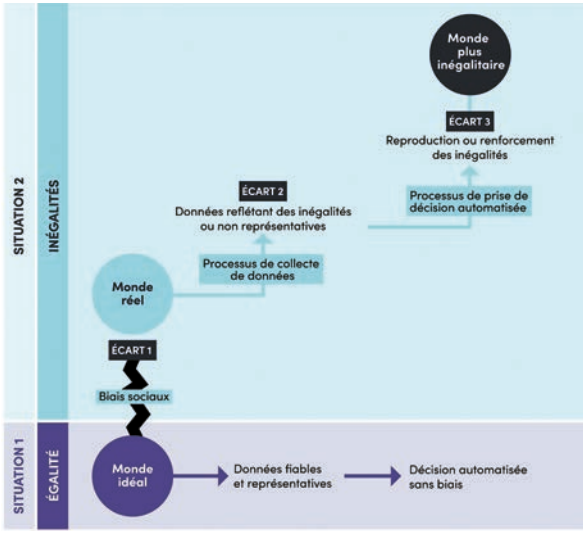
2023 – Le journal *Le Monde* publie une enquête sur les dérives de l'algorithme de la Caisse nationale des allocations familiales (CAF). Cet algorithme attribue un score aux 13,8 millions de foyers d'allocataires sur la base de critères de risque de fraude. La CAF utilise cet algorithme pour prioriser les contrôles effectués par ses agents. En cherchant des corrélations dans les données des allocataires ayant commis des erreurs ou des irrégularités dans les déclarations passées, l'algorithme a mis en évidence des profils d'allocataires susceptibles de frauder la CAF. C'est ainsi que des mères monoparentales ayant récemment déménagé et gagnant moins de 942 euros par mois étaient considérées comme plus à risque de fraudes, donc sujettes à davantage de contrôle menant à ce que certains ont dénoncé comme de la « maltraitance institutionnelle » (Geiger et collab., 2023).

Ces exemples montrent que plusieurs systèmes d'IA actuellement utilisés dans le secteur public pour assister les agents publics à accomplir leur travail sont propices à causer ou à exacerber le problème des biais algorithmiques. Pour comprendre comment émergent les biais algorithmiques, imaginons un monde dans lequel il n'y aurait aucune

4. Une traduction polie du slogan serait : « À bas l'algorithme ! »

inégalité. Dans ce monde imaginaire, l'IA serait alimentée et entraînée à partir de données représentant fidèlement cette société égalitaire. Les prédictions et les résultats produits par des algorithmes d'aide à la décision seraient précis, fiables et exempts de discrimination. Ce monde imaginaire et le fonctionnement de cette IA idéale sont représentés par la situation 1 dans la figure 1. Nous ne vivons malheureusement pas dans ce monde idéal. Le monde dans lequel nous vivons n'offre généralement pas les mêmes chances ni les mêmes occasions à tout le monde. La situation 2 dans la figure 1 plonge ses racines dans ce contexte inégalitaire. Comme nous l'observons dans de nombreux domaines, certaines personnes sont victimes de discrimination ou n'obtiennent pas les services auxquels elles ont droit en raison de leur statut socioéconomique, de leur genre ou de leur origine. On parle de « biais sociaux » pour qualifier ces biais qui existent dans la société (écart 1 dans la figure 1), sans que l'IA en soit responsable, mais que les algorithmes reproduisent ou amplifient. En effet, les données qui sont collectées dans un contexte inégalitaire ou discriminatoire vont refléter ce contexte (écart 2 dans la figure 1). C'est ainsi qu'un algorithme analysant le curriculum vitæ (CV) des personnes postulant à un emploi dans un secteur où les hommes sont majoritaires va avoir tendance à écarter les candidatures des femmes si l'algorithme est entraîné en utilisant les CV des employés actuels de l'organisation. Cet algorithme, en s'appuyant sur des données reflétant le monde réel, va reproduire et contribuer à renforcer les biais sociaux existants (écart 3 dans la figure 1).

FIGURE 1
Comment apparaissent les biais algorithmiques ?



Source : figure inspirée par Lattimore et collab. (2020).

Ainsi, si les données utilisées contiennent des biais, il y a de forts risques que les prédictions et les résultats de l'algorithme soient à leur tour biaisés. Ce phénomène se résume dans l'expression : « mauvaises données à l'entrée, résultats erronés à la sortie » (*garbage in, garbage out*). De plus, si le système algorithmique est utilisé pour prédire des risques de récidive ou de fraude, comme nous l'avons vu dans les exemples plus haut, l'algorithme peut renforcer et exacerber les préjudices envers certains groupes de la population en les soumettant à « une sur-surveillance qui confirme la prédiction [d'un comportement déviant], renforçant ainsi les contrôles dans le quartier et qui, en fin de compte, alimente encore plus les préjugés à l'égard de cette

partie de la population » (ELA, 2023, p. 12, trad.). Dans le cas des algorithmes de police prédictive, cela implique que les décisions antérieures, comme les lieux d'intervention des forces de l'ordre ou les endroits où des infractions ont été signalées, influencent les prédictions futures. Si les données d'arrestation d'un quartier sont utilisées pour réentraîner l'algorithme, la police sera systématiquement envoyée dans ce même quartier, créant un cycle où le biais initial se renforce continuellement. Avec cette boucle de rétroaction (*feedback loop*), l'algorithme énonce en quelque sorte une prophétie auto-réalisatrice contribuant à renforcer les contrôles sur un échantillon non représentatif de la population (van Giffen et collab., 2022). Cela crée un cercle vicieux où un système utilise constamment ses propres résultats pour se renforcer, amplifiant ainsi ses biais initiaux, sans tenir compte des conditions réelles sous-jacentes qui évoluent au fil du temps. Cela amène plusieurs observateurs à signaler que l'accroissement de l'utilisation de l'IA pourrait mener à un monde plus inégalitaire.

Les situations problématiques décrites dans cette section trouvent leurs origines non seulement dans la nature technique des outils algorithmiques, mais se situent aussi en grande partie dans la conception des modèles, la qualité des données utilisées et le contexte d'utilisation des systèmes d'IA sur lesquels nous nous attarderons dans le chapitre 2. En effet, au-delà de leur apparente neutralité et objectivité, les systèmes d'IA sont « influencés par des humains, des institutions et des impératifs qui orientent leurs fonctions et leur fonctionnement » (Kim Crawford, 2021, traduction citée par CSF, 2023, p. 9). Ainsi, « les systèmes décisionnels fondés sur l'apprentissage automatique ne sont pas simplement des outils technologiques. Ils font partie d'un système sociotechnique, c'est-à-dire un système dans lequel

les technologies façonnent les personnes et les organisations et sont façonnées par elles » (USAID, 2022, p. 31, trad.).

Pour mieux comprendre les préoccupations soulevées par l'utilisation des systèmes algorithmiques dans le secteur public, nous avons défini quatre enjeux qui caractérisent et structurent l'interrelation entre l'humain et la technique dans l'émergence et la gestion des biais algorithmiques. Ces enjeux sont la gouvernance algorithmique, l'acceptabilité sociale de l'utilisation de l'IA, la qualité des modèles d'IA et l'opacité des systèmes algorithmiques.

3.1 Gouvernance algorithmique

Comme nous l'avons mentionné en introduction, l'État peut incarner différents rôles lorsqu'il intervient dans le domaine de l'IA. Il peut agir 1) en tant que régulateur, 2) comme facilitateur, en définissant un cadre pour les entreprises et la population ou 3) en tant que principal acquéreur et utilisateur de solutions d'IA à son profit. Au cours des dernières années, l'accent a été mis davantage sur son rôle de régulateur et de facilitateur, plutôt que sur celui d'utilisateur de l'IA (Kuziemski et Misuraca, 2020). L'orientation actuelle de la plupart des États est donc de se concentrer sur la gouvernance *de* l'IA (les autorités publiques occupant un rôle de régulateur et facilitateur de l'introduction de l'IA au sein de la société) et moins sur la gouvernance *avec* l'IA ou gouvernance algorithmique (les autorités publiques comme consommatrices et bénéficiaires de l'IA). Dans cet ouvrage, nous nous concentrons sur ce dernier aspect : la gouvernance *avec* l'IA[5].

5. Nous verrons dans le chapitre 3 comment la gouvernance avec l'IA suscite tout de même des réflexions sur la gouvernance de l'IA dans le secteur public.

La prise de décision axée sur les données n'est pas un concept récent en administration publique. Les algorithmes modernes dans la gestion publique peuvent être vus comme une continuation des efforts de rationalisation entrepris à partir du milieu du 20e siècle décrits dans l'encadré 4. En s'appuyant sur des technologies plus avancées pour traiter de grands volumes de données et soutenir les décisions, les algorithmes associés aux techniques modernes d'apprentissage automatique permettent de traiter les données d'une manière innovante, privilégiant la prédiction à l'explication, rendant ainsi leurs compréhension et interprétation plus ardues (Levy et collab., 2021 ; Veale et Brass, 2019). De plus, les algorithmes récents s'orientent davantage vers la collecte et l'intégration des données, accentuant ainsi leur capacité à tirer profit de l'intelligence collective en matière de services publics (Vogl et collab., 2020).

ENCADRÉ 4
La quête de rationalité et d'automatisation dans la gestion publique

Le développement de mécanismes algorithmiques dans le secteur public s'inscrit dans une tradition plus ancienne de la gestion publique, qui a commencé à se former au milieu du 20e siècle. À cette époque, les gouvernements ont cherché à rendre leurs processus plus systématiques et efficaces en adoptant des approches basées sur la formalisation pour structurer les processus et la rationalisation pour optimiser les ressources. Cette tendance a atteint un point culminant dans les années 1960 avec l'introduction du Planning-Programming-Budgeting System (PPBS) aux États-Unis. Ce système de gestion intégrée a influencé de nombreux pays, qui ont adopté des méthodes de gestion publique similaires. Cet idéal décisionnel systématisé et quantifié puise son origine dans la recherche opérationnelle, qui aspire à concevoir des méthodes d'analyse neutres, scientifiques et objectives pour déterminer les actions optimales que l'État doit envisager (Harcourt, 2018). C'est également au 20e siècle

qu'ont été entreprises les premières tentatives d'automatisation à l'aide des systèmes experts, visant à coder l'expertise dans des règles logiques (Berry, 1997).

Ces ambitions de rationalisation du fonctionnement de l'État se retrouvent ensuite dans les nombreuses initiatives de modernisation ou de réforme de l'administration publique qui ont culminé avec la nouvelle gestion publique (NGP – *new public management*). Les promoteurs de la NGP mettent l'accent sur la recherche de l'efficacité et de l'efficience, l'atteinte des résultats et l'adoption de moyens favorisant l'amélioration de la performance (ex. : rémunération au mérite). L'aspiration de la NGP de transposer des principes de gestion du secteur privé vers le secteur public trouve ainsi écho dans l'adoption des systèmes de décision automatisée où la standardisation, l'automatisation et la quantification sont des composantes centrales. Ces systèmes de décision automatisée, prétendument « libérés » des biais humains, sont perçus comme prenant des décisions optimales, élevant ainsi les idéaux de la NGP à un niveau supérieur (König et Wenzelburger, 2021). Ces aspirations se retrouvent également dans le plaidoyer en faveur des décisions fondées sur des données probantes (*evidence-based policymaking*) et pointent vers une forme de gouvernance par les algorithmes qui reproduit et approfondit ce que certains nomment la gouvernance par les nombres (Newman et Mintrom, 2023 ; Paul et collab., 2024 ; Supiot, 2015).

Les outils algorithmiques modernes s'intègrent au cœur des activités administratives et favorisent l'émergence d'une gouvernance algorithmique qui façonne un nouvel ordonnancement social à l'aide de règles et de procédures informatiques complexes (Katzenbach et Ulbricht, 2019).

La gouvernance est un terme polysémique qui se définit comme « un processus d'agrégation, de coordination et de direction d'acteurs, de groupes sociaux et d'organisations, en vue d'atteindre des objectifs définis et discutés collectivement » (Le Galès, 2019, p. 298). Dans nos sociétés complexes et fragmentées, aucun acteur ne dispose des connaissances,

des ressources et de la capacité nécessaires pour gouverner seul ; l'interaction est nécessaire pour atteindre les buts. La gouvernance implique un degré minimal de stabilité, condition préalable à la coordination, pour que les acteurs développent des attentes et structurent leurs intérêts (Katzenbach et Ulbricht, 2019). Cette brève introduction au concept de gouvernance est utile pour expliquer pourquoi le terme « gouvernance algorithmique » nous semble plus approprié que celui de « régulation algorithmique » pour décrire la gouvernance par les algorithmes. En effet, la notion de « gouvernance » permet de rendre compte de la multiplicité de l'ordre social en ce qui concerne les acteurs, les mécanismes, les structures, les degrés d'institutionnalisation et la répartition de l'autorité, tandis que la notion de « régulation algorithmique » est souvent associée à l'élaboration et l'adoption de règlements, de lois ou de politiques visant à encadrer le développement et l'utilisation des algorithmes par les organisations publiques et privées.

La gouvernance algorithmique décrit généralement un « système technologique où des algorithmes informatisés sont utilisés pour collecter et analyser des données permettant de prendre des décisions qui s'appuient sur ces données. Ce système est généralement destiné à contrôler, inciter, encourager ou manipuler le comportement d'êtres humains ou celui d'autres machines » (Udoh, 2020, p. 2, trad.). La notion de « gouvernance algorithmique » permet également de comprendre comment les algorithmes structurent l'ordre social, en faisant référence à la manière dont la gouvernance est exercée par des algorithmes plutôt qu'à la gouvernance des algorithmes par des humains (Katzenbach et Ulbricht, 2019 ; Musiani, 2013). Ainsi,

> même si la gouvernance algorithmique améliore potentiellement l'efficacité des organisations publiques, l'utilisation d'outils algorithmiques implique souvent de nouvelles formes de surveillance de la population, soulève des préoccupations en matière de droits de la personne et remet en question les stratégies et les moyens de gestion des institutions publiques, ce qui peut ultimement nécessiter des changements sur les plans politique et sociétal (Pūraitė et collab., 2020, p. 2151, trad.).

De nos jours, le gouvernement par la loi tend à être renforcé – sans nécessairement être remplacé – par une gouvernance algorithmique. De plus en plus souvent, la mise en œuvre de lois et de règlements se produit en recourant à des systèmes informatisés ou à des algorithmes. Ce mouvement conduit à une individualisation de la relation à l'État. Ainsi, plutôt que de soumettre l'ensemble de la population à une loi de portée générale, les outils technologiques peuvent être employés pour orienter certaines personnes à adopter le comportement souhaité au moyen d'un guidage plus précis ou d'une programmation du comportement (Marique et Strowel, 2017). Les manifestations de la gouvernance algorithmique sont diverses et méritent une analyse approfondie pour être en mesure de saisir comment l'action humaine est de plus en plus surveillée, orientée ou même dirigée par des outils technologiques. Parfois, les algorithmes contribuent à faire respecter des normes et des règlements en repérant des comportements déviants, des pratiques frauduleuses, des infractions ou des risques de récidive de comportements problématiques. Dans d'autres situations, les algorithmes peuvent générer ou orienter une décision administrative qui va avoir des répercussions sur la vie des personnes ou des entreprises. De plus, avec l'intégration d'algorithmes dans les objets physiques, certaines actions, autrefois soumises à la décision humaine, sont désormais régies par ces systèmes d'IA. Par exemple, les algorithmes pour les véhicules

autonomes peuvent empêcher ces derniers de dépasser les limites de vitesse ou de se garer, même temporairement, dans une zone interdite. Ce sont quelques exemples des répercussions potentielles de l'IA dans la vie quotidienne (Marique et Strowel, 2017 ; Susskind, 2018 ; Veale et Brass, 2019 ; Vogl et collab., 2020).

En résumé, la gouvernance algorithmique est décrite comme une « forme d'ordonnancement social qui repose sur la coordination entre acteurs, s'appuie sur des règles et intègre des procédures épistémiques informatisées particulièrement complexes » (Katzenbach et Ulbricht, 2019, p. 2, trad.). L'intérêt pour cette forme de gouvernance s'accroît en raison des changements qu'elle entraîne sur les modes de gestion et le fonctionnement de l'État lorsqu'il intègre des algorithmes (Poussart, 2021). La plupart des observateurs estiment que la gouvernance algorithmique devient de plus en plus puissante, intrusive et envahissante en raison des nouvelles formes d'exclusion et d'asymétrie de pouvoir qu'elle engendre. Des questions d'opacité, de manque d'explicabilité et de transparence sont fréquemment soulevées et seront abordées en détail dans la suite de cet ouvrage. Cependant, certains observateurs défendent, quant à eux, l'idée qu'une gouvernance basée sur les algorithmes peut être plus inclusive, réactive et favoriser une plus grande diversité sociale (Katzenbach et Ulbricht, 2019).

Dans le secteur public, la gouvernance algorithmique a le potentiel d'améliorer l'efficacité et l'efficience des administrations. Toutefois, son déploiement soulève aussi de sérieux problèmes, tels que les préjugés, l'injustice et la discrimination dans la prise de décision. Le phénomène du biais algorithmique, où les systèmes d'IA reproduisent, perpétuent ou amplifient les biais sociaux existants, peut exacerber les inégalités et contribuer à la discrimination des

personnes vulnérables ou marginalisées, particulièrement dans le contexte actuel où les algorithmes servent principalement d'outils d'aide à la décision administrative (Alon-Barkat et Busuioc, 2023).

3.2 Acceptabilité sociale et asymétrie de pouvoir

L'acceptabilité sociale de l'utilisation des algorithmes dans les processus décisionnels n'est pas garantie et varie en fonction de la nature des actions ou décisions que les organisations publiques délèguent aux applications d'IA. Pour certains, l'utilisation d'algorithmes dans les processus décisionnels qui ont des répercussions concrètes sur les êtres humains est inacceptable. Les préoccupations à ce sujet incluent le respect de la vie privée, le manque d'équité et l'absence de supervision ou d'intervention humaine dans des décisions importantes pour les personnes concernées (De-Arteaga et collab., 2020 ; Kuziemski et Misuraca, 2020).

Bien que les responsables politiques et les dirigeants des organismes publics présentent souvent l'introduction des technologies numériques dans le secteur public comme avantageuse pour les usagers, certains observateurs et analystes estiment, quant à eux, que la collecte et l'analyse continues de données, combinées à l'utilisation de systèmes basés sur l'IA par les organisations publiques, soulèvent de nombreuses préoccupations, telles qu'un retour à l'unilatéralisme administratif et un accroissement de l'asymétrie de pouvoir entre l'État et la population. Selon eux, il est important d'examiner comment l'apprentissage automatique et la bureaucratie sont devenus des modes généralisables d'ordonnancement rationnel fondés sur l'abstraction et tirant leur autorité de prétentions à la neutralité et à l'objectivité (McQuillan, 2019). En d'autres termes, la question se pose de savoir si

l'IA facilite un rééquilibrage des pouvoirs entre l'État et les usagers ou si, au contraire, elle accroît l'asymétrie de pouvoir existante (Kuziemski et Misuraca, 2020).

Le choix du modèle d'IA retenu par les organisations publiques illustre bien ce problème d'acceptabilité et d'asymétrie de pouvoir. En effet, le modèle sélectionné et ses paramètres de fonctionnement, comme le taux de précision, ont des incidences majeures sur les résultats et les performances du système. Un certain taux d'erreurs dans les prédictions peut, par exemple, être considéré comme satisfaisant pour les employés d'une organisation publique mais inacceptable pour les usagers concernés. Ainsi, la légitimité de la prise de décision ne réside plus uniquement dans la conformité des processus décisionnels, mais découle également de l'acceptabilité des résultats produits par le SIA. Des exemples relatifs à l'acceptabilité de l'utilisation de l'IA dans le secteur public ont été présentés plus haut, lorsque la ministre néozélandaise du Développement social a mis fin à un projet visant à identifier les enfants à risque de maltraitance et de négligence ou lorsqu'un juge australien a décrété l'illégalité du système Robodebt qui visait à repérer et à recouvrer des prestations sociales indûment versées.

Plus généralement, l'acceptabilité sociale de l'utilisation de l'IA dans le secteur public suscite de nombreuses interrogations et préoccupations au sein de la population et parmi les employés de l'État. Le recours accru à l'automatisation et aux systèmes d'aide à la décision automatisée est critiqué par plusieurs en raison des risques de déshumanisation des services publics et de l'aggravation de la fracture numérique pour les personnes vulnérables.

3.3 Qualité des modèles d'IA

Les développeurs d'IA créent des modèles qui structurent le fonctionnement des algorithmes, définissant ainsi la séquence à suivre pour accomplir une tâche particulière. Ces modèles se caractérisent par des paramètres choisis par leurs concepteurs en fonction des données disponibles, des objectifs à atteindre et des capacités informatiques (DataFranca, 2024). Chaque modèle possède des caractéristiques propres et est adapté à des types de problèmes spécifiques. Ainsi, la qualité d'un modèle dépend de la qualité des données utilisées et des choix faits par les développeurs quant aux objectifs à atteindre.

Les algorithmes ne sont pas à l'abri de commettre des erreurs. Par exemple, un algorithme de détection d'objets peut mal reconnaître un objet en raison de variations d'éclairage ou de la qualité de l'image. Cette situation s'observe avec l'algorithme des autorités fiscales françaises visant à repérer les piscines non déclarées qui « aurait tendance à confondre les aires de stationnement réservées aux personnes handicapées, signalées par des marquages bleus, avec des piscines [… À l'inverse,] si la couleur de l'eau d'une piscine ne correspond pas au bon bleu, le robot ne la détecte pas » (Arène, 2024). Dans ce cas, nous constatons que l'erreur représente une inexactitude dans le résultat produit par un algorithme. Ces erreurs sont souvent causées par des limitations techniques, des erreurs de codage ou des données d'entrée incorrectes. Pour éviter ces erreurs, il est nécessaire de tester rigoureusement les modèles avant de les utiliser. Toutefois, selon le responsable d'un syndicat de la fonction publique française, plusieurs algorithmes déployés aujourd'hui dans le secteur

public « sont des prototypes qui essuient les plâtres[6], en situation réelle » (Arène, 2024).

Dans la réflexion sur la qualité, le choix du degré de précision d'un modèle est une décision importante prise lors de la création d'un modèle à laquelle les utilisateurs des systèmes algorithmiques ne prêtent parfois pas suffisamment d'attention. La précision d'un modèle (*accuracy*) exprime en pourcentage la proportion de prédictions correctes effectuées par le modèle. Un degré de précision de 100 % signifie que le modèle parvient à prédire parfaitement ce qui va se produire, c'est-à-dire qu'il ne commet aucune erreur. Bien qu'un taux de précision de 100 % soit idéal, il est rarement, voire jamais, atteint dans la pratique, en raison de la complexité des tâches à accomplir, des imperfections dans les données ainsi que des limites des algorithmes. Certaines organisations publiques utilisent des algorithmes dont le degré de précision est faible. Par exemple, le moins bon degré de précision du modèle COMPAS, utilisé pour évaluer le risque de récidive des criminels aux États-Unis, est de 65 % (Brennan et collab., 2009). Ce taux de précision signifie que COMPAS prédit correctement la probabilité qu'un criminel commette un nouveau crime s'il est libéré dans 65 % des cas. Rappelons qu'en lançant une pièce en l'air nous avons 50 % de chance de prédire le côté sur lequel elle va retomber. L'algorithme COMPAS n'est donc guère plus performant que le hasard pour prédire le risque de récidive. Étant donné le niveau de précision du modèle, il convient de se demander s'il est raisonnable de l'utiliser pour éclairer une décision qui a des conséquences majeures sur la vie des personnes concernées.

6. L'expression « essuyer les plâtres » s'utilise pour parler de personnes qui sont les premières à expérimenter un projet, un produit ou une situation et qui doivent en supporter les défauts ou les ajustements nécessaires.

De plus, les systèmes d'apprentissage automatique appliquent souvent un modèle statistique uniforme à toutes les décisions, ce qui peut entraver l'examen des spécificités d'un cas particulier et s'avérer inapproprié pour les décisions nécessitant une analyse nuancée de la situation et l'exercice d'un jugement flexible. En effet, en analysant de grands ensembles de données, les algorithmes privilégient les groupes majoritaires et diluent la voix des groupes minoritaires. Pour corriger ce problème, il est possible de concevoir un modèle qui distinguerait les groupes en créant des sous-ensembles de données (*cluster*) homogènes en fonction de caractéristiques distinctives. Cependant, dans plusieurs cas, il peut être difficile de prédire le comportement d'un individu en se basant uniquement sur les informations du groupe auquel il appartient (Cobbe, 2019). Ce problème a par exemple été observé dans le secteur de l'éducation, où des algorithmes visant la prédiction du décrochage scolaire, la notation automatisée des examens et la gestion des admissions aux études supérieures n'atteignent pas leur objectif et risquent de générer des inégalités lorsque les modèles sont généralisés sur l'ensemble de la population étudiante (Baker et Hawn, 2022). Dans le domaine de l'aide internationale, les modèles d'évaluation de la solvabilité des emprunteurs, utilisés dans les programmes de microcrédit, lorsqu'ils sont appliqués de manière identique aux hommes et aux femmes (modèles « agnostiques »), tendent à sous-estimer la capacité de remboursement des femmes comparativement à un modèle de crédit où chaque genre est évalué séparément. Ainsi, en créant des modèles distincts pour chaque genre, un plus grand nombre de femmes se voient accorder un prêt, y compris celles qui auraient été injustement refusées par le modèle agnostique, malgré une capacité de remboursement satisfaisante (USAID, 2022). Les particularités et les

paramètres des modèles peuvent donc engendrer des décisions biaisées, affectant négativement certaines personnes ou certains groupes et exacerber les inégalités sociales existantes.

Pour plusieurs observateurs, le problème ne réside pas seulement dans le fait que les algorithmes reflètent principalement les préjugés existants dans notre société, mais également dans le fait que les algorithmes amplifient considérablement ces biais, sans que s'exerce un contrôle adéquat pour les éviter (Panch et collab., 2019). En effet, les ensembles de données utilisés pour l'apprentissage automatique sont créés par des humains. Bien que les modèles soient paramétrés pour respecter un certain taux d'erreur, ils ne sont pas toujours évalués ou testés en tenant compte de tous les scénarios possibles. Ainsi, lorsqu'ils sont déployés au sein des organisations publiques, des anomalies et des erreurs non détectées lors de la conception du modèle peuvent engendrer des décisions erronées qui suscitent des répercussions négatives pour des personnes ou des communautés déjà parfois marginalisées (Cobbe, 2019). Nous reviendrons sur ce problème dans le prochain chapitre en décrivant le biais de déploiement.

3.4 Opacité des systèmes algorithmiques

Les prédictions et les résultats produits par les algorithmes sont parfois tellement insaisissables que même leurs développeurs sont incapables d'expliquer précisément comment un algorithme en est arrivé à produire un résultat donné[7]. Le manque d'explicabilité et de transparence dans

7. Ces enjeux existaient déjà avec les systèmes d'intelligence artificielle symboliques, mais ils étaient plus aisément traçables, puisque la programmation de ces systèmes plus simples était facilement vérifiable. « L'intelligence artificielle symbolique désigne l'ensemble des approches et techniques en intelligence artificielle qui sont fondées sur des représentations "symboliques" (c'est-à-dire lisibles par l'homme).

certains processus décisionnels algorithmiques ainsi que la difficulté de contester ces décisions peuvent ébranler la confiance de la population et porter atteinte à la légitimité et à la responsabilité des institutions publiques, tout en menaçant le respect de la vie privée et les droits fondamentaux (Diakopoulos, 2015 ; Busuioc, 2021). De plus, le manque d'explicabilité contrevient à l'obligation de motiver les actes administratifs qui a vu le jour dans la seconde moitié du 20e siècle pour atténuer l'unilatéralisme administratif. Cette exigence permet aux usagers de comprendre les raisons justifiant les décisions administratives et facilite une éventuelle contestation de la décision en fonction des voies de recours existantes (Autin, 2011).

L'opacité des systèmes algorithmiques représente un problème majeur qui affecte l'ensemble du cycle de l'apprentissage automatique, et plus particulièrement celui du déploiement d'un système et de la gestion de ses répercussions sur la vie des personnes ou sur la société en général. Cédric Villani, mathématicien et auteur d'un rapport sur l'IA remis au premier ministre français, indiquait d'ailleurs à ce sujet que, « [d]ans un contexte où l'IA est susceptible de reproduire des biais et des discriminations, et à mesure de son irruption dans nos vies sociales et économiques, être en mesure "d'ouvrir les boîtes noires" tient de l'enjeu démocratique » (Villani, 2018, p. 21). L'expression « boîte noire » est ainsi utilisée pour décrire la nature opaque de l'apprentissage automatique et, plus encore, de l'apprentissage profond. Elle fait référence au processus qui consiste à transformer des données du monde réel (par exemple, les pixels d'une

[…] L'IA symbolique fut le paradigme dominant de la recherche en IA depuis son origine au milieu des années 50 jusqu'au début des années 90 » (DataFranca, 2024).

radiographie) à travers plusieurs étapes appelées « couches ». Chaque couche prend en entrée les informations issues de la couche précédente et les transforme progressivement, de manière à extraire des caractéristiques de plus en plus complexes. Par exemple, les données d'entrée, comme les pixels d'une radiographie, sont transformées par une première couche du modèle qui en extrait des caractéristiques de base, telles que des contours ou des bords simples, qui permettent de reconnaître des formes élémentaires, comme des lignes ou des points lumineux. Ensuite, les couches suivantes analysent ces informations pour repérer des motifs plus sophistiqués, comme des anomalies spécifiques (comme des types de lésions), jusqu'à produire un résultat final, appelé « sortie » (*output*) dans le jargon de l'IA, tel qu'une prédiction ou un diagnostic. Par exemple, pour « des patients présentant des métastases hépatiques, cette prédiction peut inclure la prédiction de la réponse à la chimiothérapie, le pronostic d'une maladie sans récidive pour les patients traités ou les probabilités de survie » (Montagnon et collab., 2020, p. 4, trad.).

En pratique, il peut y avoir des centaines de couches et l'influence de chaque élément d'une couche est ajustée en fonction des paramètres du modèle. Dans cet exemple, il est pratiquement impossible de comprendre exactement quelles caractéristiques l'algorithme a utilisées pour établir le diagnostic. Ce processus permet d'aboutir à des résultats certes impressionnants, mais qui sont difficiles à expliquer. Par conséquent, ni les scientifiques des données développant un algorithme ni les utilisateurs comme les médecins, et encore moins les usagers (les patients, dans notre exemple), ne sont capables de savoir comment l'algorithme a produit un résultat ou une prédiction particulière (Panch et collab., 2019).

Il existe plusieurs formes d'opacité algorithmique qui rendent difficiles la compréhension, le suivi ou l'interprétation des processus internes des algorithmes.

1. L'opacité intentionnelle fait référence à la dissimulation volontaire des mécanismes internes d'un système, souvent pour protéger des secrets commerciaux ou des droits de propriété intellectuelle. Cela signifie que les entreprises ou les développeurs qui conçoivent des algorithmes choisissent de ne pas divulguer certains aspects techniques de leur fonctionnement, afin de préserver leur avantage concurrentiel ou d'empêcher la reproduction de leur technologie par d'autres. Par conséquent, même des personnes ayant les compétences techniques ne sont pas en mesure de comprendre ou d'analyser ces systèmes, car les informations nécessaires pour y parvenir sont délibérément cachées. Cela contribue à rendre les algorithmes opaques pour le grand public et même pour les experts.
2. L'opacité en tant qu'analphabétisme technique désigne une situation où seule une minorité de personnes ayant une expertise avancée peuvent comprendre comment fonctionne un système algorithmique. Cela signifie que la majorité des utilisateurs, qui ne possèdent pas cette connaissance technique, ne sont pas en mesure de comprendre ou d'analyser le fonctionnement interne des algorithmes. Cet écart de compétence crée une dépendance à l'égard des experts et contribue à l'opacité, car ceux qui ne maîtrisent pas le langage ou les concepts techniques ne peuvent pas évaluer ni remettre en question les décisions prises par ces systèmes.
3. L'opacité intrinsèque fait référence à la complexité inhérente à certains algorithmes, en particulier dans les systèmes d'apprentissage profond (*deep learning*). Ces systèmes reposent sur des réseaux neuronaux avec de multiples couches et des millions de paramètres

> interagissant entre eux. Cette structure rend les processus décisionnels extrêmement difficiles à interpréter, même pour les concepteurs du système. Cette opacité résulte donc de l'impossibilité humaine à déchiffrer ou expliquer des processus qui dépassent les capacités d'analyse traditionnelles (Burrell, 2016).

Ces formes d'opacité peuvent se superposer, rendant encore plus difficile la compréhension du fonctionnement de l'algorithme. En conséquence, les agents publics se retrouvent dans l'incapacité de vérifier et d'expliquer leurs décisions aux usagers, les privant ainsi de la possibilité de demander une révision ou de contester les décisions basées sur les résultats produits par l'algorithme (Cobbe, 2019).

Les enjeux d'opacité et de responsabilité dans les processus décisionnels ne sont pas nouveaux et existaient déjà à l'ère prénumérique. La littérature sur la responsabilité des décideurs publics mentionne qu'un certain degré d'opacité est inhérent à la prise de décision administrative. Toutefois, contrairement aux processus décisionnels impliquant uniquement des humains, qui demeurent largement opaques, les systèmes d'aide à la décision automatisée peuvent, dans certains cas, devenir transparents et être soumis à des audits (König et Wenzelburger, 2021).

En somme, les inquiétudes liées au déploiement d'algorithmes d'apprentissage automatique dans les organisations publiques sont exacerbées par leur opacité et suscitent des préoccupations quant à leur transparence, leur compréhension et leur explicabilité. Des interrogations juridiques émergent également au sujet de la légalité des actions entreprises sur la base d'algorithmes, en particulier en matière de compétence, de motivation et de responsabilité administrative en cas d'erreur ou de dysfonctionnement

algorithmique (Cluzel-Métayer, 2018). De plus, l'opacité algorithmique sape les bases du débat contradictoire, car elle empêche les parties d'examiner et de contester les éléments influençant une décision. L'opacité compromet également la transparence administrative, qui est une condition essentielle pour que les usagers puissent comprendre les processus décisionnels qui les concernent. Pour toutes ces raisons, plusieurs spécialistes incitent les agents publics à la prudence et à la vigilance face aux « nouveaux pièges des algorithmes [qui requièrent une certaine] distance critique » (Marique et Strowel, 2017, p. 521).

CHAPITRE 2

Décryptons les biais algorithmiques

Comme nous l'avons mentionné dans le chapitre précédent, les organisations publiques utilisent l'intelligence artificielle pour améliorer la qualité des services offerts à la population et aux entreprises, tout en réduisant les délais et les coûts d'exploitation. L'utilisation croissante de l'IA par les organisations publiques suscite néanmoins de nombreuses préoccupations relatives aux répercussions et aux problèmes engendrés par l'intégration des algorithmes dans les processus de décision administrative. Pour certains, ces problèmes représentent des impasses insurmontables, car ils sont profondément enracinés dans la nature des systèmes algorithmiques ainsi que dans les structures de pouvoir en place et les modèles économiques dominants (Eubanks, 2018 ; Zuboff, 2020). D'autres, en revanche, y voient des problèmes techniques liés à une technologie embryonnaire et croient qu'il sera possible de les résoudre avec le temps, à mesure que la technologie évoluera et mûrira.

Dans le contexte actuel, où les préoccupations liées à la discrimination et à l'objectivité des décisions occupent une place centrale dans la société, l'utilisation des algorithmes est présentée comme une solution prometteuse. En effet, les algorithmes sont généralement perçus comme des outils dotés d'une forme de neutralité et de capacités d'objectivité

supérieures à celles des humains (Cluzel-Métayer, 2020). Des chercheurs mettent d'ailleurs en garde contre une opposition trop simpliste entre une gouvernance générale par loi et une gouvernance individualisée par les algorithmes. Ces chercheurs soulignent que le recours à des algorithmes permet de personnaliser les services offerts aux usagers, donc de favoriser l'équité qui est parfois absente d'une application générale de la réglementation qui est insensible et aveugle aux différences existantes au sein de la population (Marique et Strowel, 2017). À titre d'illustration, l'application Parcoursup a suscité une vive polémique, en France, lorsque le grand public a découvert que les décisions déterminant l'avenir de centaines de milliers de bacheliers français étaient prises à l'aide d'un algorithme dont l'objectif était d'éliminer l'opacité du processus de répartition des étudiantes et des étudiants (Gribonval, 2021). En effet, auparavant, ces décisions étaient prises par les responsables de la gestion des admissions des établissements d'enseignement supérieur et suscitaient des critiques en raison de l'opacité du processus et des iniquités alléguées dans la sélection des candidatures (Cluzel-Métayer, 2018). Ces inquiétudes étaient révélatrices d'un manque de transparence, de neutralité et d'objectivité du processus traditionnel que les concepteurs de l'algorithme souhaitaient corriger à l'aide d'une solution technique.

Il est toutefois important de ne pas surestimer l'apparente neutralité de ces systèmes. Bien que l'objectivité algorithmique soit attrayante, elle peut s'avérer illusoire pour deux raisons. Premièrement, la conception de la plupart des algorithmes repose sur des décisions prises par les autorités compétentes au sein des organisations publiques. Les choix des tâches que l'algorithme doit accomplir ainsi que les nombreuses décisions relatives à ses paramètres de fonctionnement et aux données qu'il utilise sont pris par

des humains et vont avoir des répercussions importantes sur ses performances et sur les personnes concernées par ces décisions algorithmiques. Deuxièmement, ces algorithmes s'appuient sur de vastes ensembles de données et ils risquent de perpétuer, voire d'amplifier, les biais préexistants, contenus dans ces ensembles de données, comme nous l'avons vu dans la figure 1 représentant l'origine des biais algorithmiques. C'est pour ces raisons que des organismes de protection des droits, comme la Commission nationale de l'informatique et des libertés (CNIL, 2023), en France, ont appelé à une vigilance accrue en ce qui concerne cette prétendue objectivité des algorithmes d'aide à la décision, dans un contexte où leur intégration s'accélère dans le secteur public (Cluzel-Métayer, 2018).

Rappelons que les algorithmes d'IA sont des ensembles de procédures mathématiques et statistiques appliquées à des données, dont la conception et la validation requièrent des compétences spécialisées. Des erreurs peuvent être faites et se produire à chacune des étapes de leur développement et de leur utilisation, telles que la définition des objectifs, l'encodage des données, la détermination du taux de précision des décisions et prédictions, la représentativité du contexte d'application de l'algorithme avec celui dans lequel il a été conçu, l'interprétation des résultats par les concepteurs, les usagers, les employés ou les responsables des organisations publiques (Waller et Waller, 2020).

1. QU'EST-CE QU'UN BIAIS ALGORITHMIQUE ?

Dans le langage courant, le terme « biais » est souvent empreint d'une connotation négative. La notion est plus ambiguë lorsqu'il est question de « biais algorithmique ». Dans son acception la plus neutre, le biais algorithmique

désigne un écart systématique dans les résultats (*output*) ou la performance d'un algorithme, par rapport à une norme ou à un standard prédéfini (Antony, 2016 ; Danks et London, 2017 ; Johnson, 2021). Cela signifie qu'il y a un écart, une erreur statistique, entre le résultat produit par l'algorithme et le résultat attendu. D'un autre côté, un algorithme peut présenter un biais moral ou social en fonction de la norme retenue dans son fonctionnement (Fazelpour et Danks, 2021). Certains auteurs préfèrent dans ce cas utiliser le terme « injuste » plutôt que « biaisé », réservant ce dernier pour sa signification statistique et privilégiant les termes « juste » ou « injuste » pour qualifier les enjeux sociaux ou moraux de l'utilisation des algorithmes (Baker et Hawn, 2022). Dans cet ouvrage, bien que nous utilisions principalement le terme « biais algorithmique » dans son sens le plus neutre, nous abordons également les cas problématiques d'un point de vue moral et social. Ainsi, nous utilisons le terme « biaiser » pour qualifier un comportement discriminatoire dans la société, plutôt que la définition statistique plus restrictive. Nous utiliserons parfois le terme « discriminer » pour faire directement référence aux notions d'injustice et d'iniquité sociétale (Friedler et collab., 2016).

La littérature en sciences sociales qui porte sur les algorithmes dans le secteur public se concentre principalement sur les enjeux politiques, juridiques, économiques et sociaux des systèmes algorithmiques, plutôt que sur leurs dimensions informatiques et techniques (Levy et collab., 2021). Nous adoptons une approche similaire, car, même s'il est essentiel de prendre en compte les aspects informatiques et techniques de l'IA, l'intégration de systèmes algorithmiques dans les structures institutionnelles met en évidence l'importance du facteur humain dans la sélection, la création, la mise en

œuvre et l'utilisation de ces outils. Cette approche permet également de souligner l'influence et la contribution des humains dans l'amplification, l'atténuation ou la résolution des problèmes relatifs aux biais algorithmiques. En résumé, « le biais algorithmique ne réside pas seulement dans le code informatique ou les mathématiques, mais il dépend également des domaines d'application, des objectifs d'utilisation de l'algorithme et d'autres facteurs contextuels » (Fazelpour et Danks, 2021, p. 2, trad.).

2. QUELS SONT LES PRINCIPAUX TYPES DE BIAIS ALGORITHMIQUES ?

Dans les pages suivantes, nous décrirons différents types de biais algorithmiques en fonction du moment où ils apparaissent dans le pipeline de l'IA. Ce pipeline comprend principalement les étapes techniques du cycle de développement d'un système d'IA, telles que la production ou la collecte des données, la création des modèles ainsi que l'évaluation et le déploiement du système (Suresh et Guttag, 2021). Toutefois, comme nous abordons les biais algorithmiques dans une perspective plus large, nous élargirons ce cadre en y intégrant des dimensions sociales. Cela signifie que nous ajouterons des étapes préliminaires concernant la définition du problème à résoudre par l'IA (À quel besoin répond l'IA ?), le choix des outils d'apprentissage automatique et la sélection de leurs fonctionnalités (Quel objectif doit atteindre l'algorithme ?). Ces moments sont propices à de nombreuses décisions, parfois politiques, susceptibles d'introduire des biais. En effet, les analystes de politiques publiques ont depuis longtemps montré que la manière de définir un problème public influence directement les

solutions envisageables et les solutions choisies pour le résoudre. Il en va de même pour l'IA où la définition du problème à résoudre influence grandement les choix en matière de sélection et de développement ou d'adaptation des solutions d'IA retenues.

Pour différencier les catégories des biais les plus courants, il existe plusieurs taxonomies et classifications, souvent basées sur cette approche séquentielle du cycle de développement et d'utilisation de l'IA (Barocas et collab., 2022 ; Hellström et collab., 2020 ; Mehrabi et collab., 2022 ; Silva et Kenney, 2019 ; Suresh et Guttag, 2021 ; Mitchell et collab., 2021). Les pionniers dans cet exercice de classification ont jeté les bases d'un cadrage séquentiel des biais, en les catégorisant en fonction du contexte dans lequel ils apparaissent, qu'il s'agisse des biais historiques ou sociaux préexistants, de limitations techniques ou des conséquences du déploiement et de l'utilisation des outils algorithmiques par les utilisateurs (Friedman et Nissenbaum, 1996). Comme le montrent ces travaux précurseurs, une grande partie des biais détectés dans les systèmes algorithmiques trouvent leur origine en dehors du développement à proprement parler des modèles, qu'il s'agisse de biais historiques préexistants, d'enjeux associés à la mesure ou à la définition des indicateurs, de la collecte des données ou de l'utilisation des résultats générés par le modèle (Baker et Hawn, 2022).

Nous adoptons une approche similaire pour classer les types de biais algorithmiques, en parcourant le pipeline de l'IA et en repérant les moments où ces biais apparaissent. Certains biais peuvent émerger à plusieurs étapes du pipeline ou se manifester lors d'une seule étape du cycle de développement, de déploiement ou d'utilisation de l'IA et avoir tout

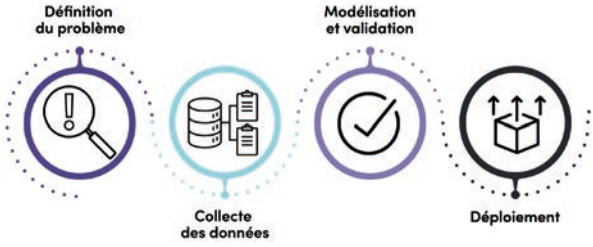
de même des répercussions importantes plus tard dans le processus. Nous reconnaissons également que certains biais peuvent provenir de facteurs externes, tels que les préférences, les intérêts ou les valeurs des personnes ou des organisations qui contribuent au développement des algorithmes. Ainsi, nous adopterons une vision globale, considérant tout le processus de mise en place d'un système d'aide à la décision, depuis la formulation du besoin d'intégration de l'IA dans une organisation jusqu'à l'évaluation des effets concrets de son utilisation sur les parties prenantes concernées[1].

Notre objectif n'étant pas de répertorier les nombreuses taxonomies et classifications des biais algorithmique existantes, nous avons opté pour l'approche de Fazelpour et Danks (2021), qui offre une précision suffisante pour reconnaître les différentes étapes du cycle de vie de l'IA, tout en étant suffisamment large pour y ajouter des éléments issus d'autres taxonomies. La figure 2 illustre le pipeline de l'IA dans lequel s'insèrent quatre grandes catégories de biais :

1. les biais dans la définition du problème,
2. les biais dans les données,
3. les biais dans la modélisation et la validation,
4. les biais dans le déploiement.

1. Une dernière étape existe lorsqu'un algorithme n'est plus utilisé. Plusieurs actions doivent être entreprises pour le désactiver correctement. Les données et les résultats générés doivent être sauvegardés conformément aux lois et règlements sur la protection des renseignements personnels. Il est également nécessaire d'archiver toute la documentation sur le fonctionnement de l'algorithme pour d'éventuels contrôles ou audits. Finalement, l'algorithme doit être retiré du système d'IA où il était actif, ce qui peut nécessiter la suppression de son code ou la reconfiguration du système afin d'assurer son bon fonctionnement sans l'algorithme ou pour intégrer un nouvel algorithme en remplacement.

FIGURE 2
Pipeline de l'IA



Source : figure inspirée de Fazelpour et Danks (2021).

2.1 Biais dans la définition du problème

La première étape de la conception d'un algorithme consiste à définir le problème qu'il doit résoudre. Il s'agit de préciser l'objectif de l'algorithme et d'expliquer la raison de son développement. Des recherches récentes soulignent que les objectifs de ceux qui souhaitent utiliser l'IA sont parfois mal définis, qu'ils sont incompatibles avec les capacités de prédiction ou de décision d'un système algorithmique ou que les organismes publics souhaitant adopter des outils d'IA ne disposent pas toujours de la maturité numérique requise pour en assurer une mise en œuvre efficace (Dobbe et collab., 2018 ; Eckhouse et collab., 2019 ; Silva et Kenney, 2019 ; Selbst et collab., 2019).

Pour surmonter ces problèmes, une organisation publique devrait réfléchir aux buts globaux en intégrant l'IA, à l'étendue de sa marge de manœuvre pour atteindre ces buts ainsi qu'à la façon dont le système algorithmique contribuera à leur réalisation (Mitchell et collab., 2021). Il convient également d'analyser comment l'algorithme va s'intégrer dans les processus administratifs ou dans la structure organisationnelle

dont il fera bientôt partie. En effet, le choix de recourir à des algorithmes prédictifs pour résoudre des problèmes complexes peut affecter les dynamiques institutionnelles (Aizenberg et van den Hoven, 2020) et compromettre des valeurs importantes pour la société. Il est donc essentiel de mener des réflexions préalables sur l'acceptabilité et les impacts sociétaux des systèmes d'IA utilisés par les organisations publiques. Ces discussions devraient inclure divers acteurs, tels que les responsables politiques, les représentants de la société civile, les syndicats et les chercheurs de diverses disciplines, afin de participer aux décisions, d'informer et de guider les débats, notamment sur les conditions dans lesquelles ces changements sont moralement ou socialement acceptables (Fazelpour et Danks, 2021).

Les préoccupations liées à un modèle algorithmique jugé biaisé peuvent dans certains cas résulter de choix établis préalablement lors de l'élaboration de la politique dont l'algorithme va venir soutenir la mise en œuvre. Lors de l'élaboration des politiques publiques, les parties prenantes impliquées ont souvent des objectifs, des valeurs et des intérêts divergents, voire contradictoires. Cela rend difficile l'adoption d'une politique qui est perçue comme acceptable et juste par tout le monde. Dans ce contexte, un modèle algorithmique, même techniquement équitable, peut être critiqué si ses résultats soutiennent des objectifs auxquels certains acteurs s'opposent. En d'autres termes, la diversité des perspectives observée lors de la phase d'élaboration d'une politique complique la conception de solutions technologiques consensuelles, car des désaccords sur les objectifs fondamentaux d'une politique ne peuvent être simplement résolus par des modifications techniques, telles que l'ajout de données ou des choix de modélisation différents. Si l'on est en désaccord avec l'objectif global d'une politique, un

modèle qui fait avancer cet objectif ne sera pas acceptable, même s'il est juste et équitable eu égard à ladite politique (Mitchell et collab., 2021). À cet égard, nous avons vu que plusieurs algorithmes sont utilisés pour établir des scores de risque (par exemple, le risque de décrochage scolaire, de fraude ou de récidive) afin de guider les interventions et les décisions des agents publics. Cependant, l'utilisation de grands ensembles de données pour prendre des décisions individuelles soulève des questions. Certains critiquent cette approche et remettent en question le fondement de l'utilisation de grands ensembles de données pour régler des cas individuels. Selon eux, utiliser des statistiques de groupe pour prédire le comportement ou le futur d'un individu pourrait mener à des décisions biaisées ou injustes, car les particularités de chaque personne ne sont pas nécessairement prises en compte dans ces analyses (Eckhouse et collab., 2019).

Dans ce contexte, une grande partie de la discussion technique sur l'équité algorithmique postule l'existence d'un consensus au sujet des objectifs du modèle et l'acceptation du groupe de personnes affectées par la décision algorithmique. Cependant, chacun de ces aspects implique des choix importants qui, bien qu'ils puissent être définis par des responsables politiques ou des acteurs externes au processus de construction de l'algorithme, sont essentiels pour déterminer si le modèle contribuera à promouvoir l'équité dans la société (Mitchell et collab., 2021). Dans la plupart des cas, les responsables politiques poursuivent des objectifs complexes, parfois contestés ou volontairement flous. Chaque orientation, décision ou aspect lié à l'utilisation et aux objectifs d'un système algorithmique doit être réfléchi en tenant compte des valeurs et des normes partagées qui peuvent elles-mêmes introduire des biais potentiels.

Des biais peuvent ainsi émerger lorsque la définition d'un problème ou le choix des variables cibles ne parviennent pas à refléter les objectifs complexes du monde réel (Mitchell et collab., 2021). À ce stade, il ne s'agit pas d'erreurs techniques liées au modèle ou au choix des variables, mais plutôt de lacunes dans la définition et la formulation des objectifs avant leur intégration dans un modèle algorithmique. C'est par exemple le cas du **biais d'omission du gain** qui survient lorsque des objectifs clés (ou des sources potentielles de bénéfices) sont ignorés dans les variables cibles. Cela est particulièrement courant lorsque l'on tente de simplifier un problème complexe en vue d'agir sur une partie de celui-ci (Kleinberg, Ludwig et collab., 2018). Des biais peuvent aussi survenir si la définition du problème conduit l'algorithme à traiter des groupes différents de manière identique. Cela se produit lorsque la diversité des personnes concernées est négligée lors de cette première étape du pipeline de l'IA. Cette simplification excessive peut entraîner un **biais d'agrégation**, où les caractéristiques spécifiques ou les besoins de sous-groupes sont ignorés (Suresh et Guttag, 2021). Par exemple, si une université définit la « réussite des étudiants » de manière trop large, sans tenir compte des différences dans les défis auxquels font face certains groupes, tels que les étudiantes et étudiants de première génération ou appartenant à des minorités, l'algorithme risque de traiter ces groupes distincts comme une population homogène, en négligeant des facteurs importants à l'atteinte des objectifs. Ainsi, des recherches sur un algorithme de détection des émotions dans la population étudiante ont montré que l'algorithme, entraîné sur un ensemble de données combinant des étudiantes et des étudiants provenant de milieux urbains, ruraux et suburbains, est moins performant pour chacun de ces groupes que des algorithmes entraînés uniquement sur

des données propres à chaque sous-groupe (Baker et Hawn, 2022 ; Ocumpaugh et collab., 2014).

Lors des premières étapes du développement d'un modèle d'IA, les développeurs peuvent prendre ou influencer des décisions concernant les variables cibles, ce qui peut avoir des répercussions sur plusieurs parties prenantes. Par exemple, pour l'attribution de logements à des personnes sans-abri, on pourrait décider de prédire quels individus risquent de devenir des utilisateurs à forte demande de services publics au cours de l'année, en les priorisant pour l'octroi d'un logement afin de réduire les dépenses publiques. Utiliser le risque des répercussions budgétaires comme critère de décision pour cet outil est sans doute moralement et politiquement acceptable, puisque les économies réalisées grâce à l'octroi d'un logement pourraient être réinvesties dans d'autres programmes ou pour d'autres personnes. Cependant, cet outil pourrait accorder une moindre priorité aux jeunes, qui induisent moins de dépenses immédiates, mais qui risquent d'entraîner des coûts plus importants et des souffrances accrues à long terme. Bien que les concepteurs puissent percevoir cet outil de triage comme juste et objectif, il repose néanmoins sur des choix et des décisions quant à ce qui doit être prédit et quant aux critères utilisés pour générer ces prédictions. Si ces décisions ne sont pas clairement formulées avant la conception du modèle, les développeurs pourraient intégrer implicitement une orientation politique au modèle en choisissant eux-mêmes les variables cibles en fonction de leur propre compréhension ou préférences (Levy et collab., 2021). Cette approche pourrait ainsi générer des bénéfices immédiats pour le groupe priorisé tout en ayant des conséquences négatives à long terme pour d'autres groupes, notamment les plus vulnérables (Obermeyer et collab., 2019).

C'est également lors de la définition du problème que se posent des questions relatives aux moyens à mettre en œuvre pour atteindre les objectifs, ainsi que des aspects techniques, tels que la détermination de ce qui sera considéré comme un résultat acceptable, la gestion des résultats positifs ou négatifs et la distinction entre les faux positifs et les faux négatifs. Pour comprendre ces notions, prenons l'exemple du microcrédit, qui permet notamment de lutter contre la pauvreté. Le microcrédit permet à des populations marginalisées, bien souvent des femmes, d'avoir accès à des prêts sans garantie pour démarrer des activités génératrices de revenus. Comme pour n'importe quel prêt, les institutions de microcrédit cherchent à évaluer la capacité des personnes qui empruntent à rembourser leurs prêts pour assurer la viabilité des programmes de microcrédit. Pour ce faire, les institutions financières créent des modèles de notation de crédit qui analysent divers facteurs, tels que les données financières et personnelles, afin de prédire le risque de défaut de paiement et de prendre des décisions éclairées sur l'octroi des prêts. L'intelligence artificielle est de plus en plus utilisée dans ces modèles de notation de crédit pour analyser des volumes massifs de données et affiner les prédictions de risque. Dans ce cas, des questions se posent sur les objectifs du microcrédit qui vont affecter la manière dont les facteurs de risque, tels que la solvabilité, vont être définis. De plus, des choix importants doivent être faits pour paramétrer le modèle en matière de quantité et de type d'erreurs acceptables. Il s'agit ici de la marge d'erreur et des faux positifs ou faux négatifs. Si le modèle d'IA commet des erreurs en identifiant à tort certaines personnes comme à haut risque (faux négatifs), des candidates, en particulier celles qui sont issues de communautés marginalisées, pourraient être exclues de l'accès au microcrédit. Cela peut limiter leurs possibilités

d'entreprendre ou de créer des activités génératrices de revenus, prolongeant ainsi les inégalités économiques. D'un autre côté, les faux positifs où des personnes considérées comme à faible risque se révèlent en réalité incapables de rembourser leurs prêts peuvent affecter les institutions de microcrédit en augmentant les taux de défaillance et, éventuellement, remettre en question leur viabilité. Cette situation peut pousser ces institutions à durcir leurs critères d'accès au financement, réduisant ainsi les chances pour d'autres personnes d'obtenir des prêts à l'avenir (USAID, 2022).

Avec l'exemple du microcrédit, nous observons que la manière dont une situation est présentée ou dont un problème est formulé influence les décisions et les résultats d'un modèle d'IA et risque d'entraîner un **biais de cadrage**. Le biais de cadrage se produit lorsque les modèles d'IA ne prennent en compte que certaines parties du problème en fonction des informations fournies ou des préférences de certaines parties prenantes. Ce cadrage peut influencer les résultats et conduire à des décisions erronées ou injustes.

La controverse autour de l'algorithme COMPAS, utilisé par les services correctionnels aux États-Unis pour évaluer le risque de récidive et guider les décisions des juges en matière de libération conditionnelle ou de détermination des peines, a mis en lumière un biais de cadrage. À la suite des révélations de *ProPublica* sur les biais raciaux de cet algorithme, les débats se sont concentrés sur la question de l'équité, un concept qui peut être défini de différentes façons (comme l'égalité des chances, l'égalité des occasions ou la parité prédictive). Selon Northpointe, l'entreprise ayant conçu COMPAS, les scores de risque respectaient les normes d'équité en matière de parité de prédiction : la proportion de prédictions correctes parmi les individus identifiés comme à risque était similaire entre tous les groupes, y compris les

groupes protégés (comme les groupes raciaux ou de genre). Concrètement, Northpointe affirmait que, si l'algorithme prédisait qu'une personne allait récidiver, la probabilité que cette prédiction soit exacte était la même, que l'individu soit de couleur ou blanc. Cependant, les détracteurs de COMPAS soulignaient que l'algorithme présentait un biais à l'encontre des personnes de couleur, qui étaient associées à un taux de faux positifs plus élevé. Ils arguaient que cet outil compromettait les principes d'égalité des chances et d'égalité des occasions (Larson et collab., 2016). Ainsi, étant donné la diversité des acceptions du terme « équité », un biais peut émerger selon la manière dont les mesures de succès et le problème sont définis (Srinivasan et Chander, 2021). Cela illustre comment différents critères d'équité peuvent mener à des conclusions divergentes quant au caractère biaisé ou non d'un système algorithmique.

En conclusion, des objectifs mal définis créent un terrain propice à la propagation de biais dans les étapes ultérieures du pipeline de l'IA. Comme le dit si bien la maxime, « mal nommer un objet, c'est ajouter au malheur de ce monde ». Ainsi, une mauvaise définition du problème peut entraîner des conséquences importantes sur la performance de l'algorithme.

2.2 Biais dans les données

Afin de créer des modèles capables de faire des prédictions et d'atteindre les objectifs, la démarche habituelle dans le développement de modèles d'IA consiste à entraîner un algorithme sur un ensemble de données.

> Les données constituent la matière première de l'IA, […] elles en sont le carburant. […] Celles-ci se trouvent, par exemple, sur les réseaux sociaux, dans des dossiers médicaux […] ou encore dans l'historique d'utilisation de moteurs de recherche ou d'outils de

> géolocalisation. Elles peuvent concerner les comportements des individus, leurs mouvements ou localisations, de même que leurs attributs physiologiques (CSF, 2023, p. 3).

À cette étape, la qualité des données est fondamentale pour le bon fonctionnement des modèles d'IA. Des nutritionnistes disent que nous sommes ce que nous mangeons. En appliquant ce principe aux algorithmes, nous constatons qu'ils reflètent la qualité des données sur lesquelles ils sont entraînés. Ainsi, des données biaisées ou de mauvaise qualité peuvent entraîner des modèles qui produisent des résultats incorrects, des décisions injustes ou des prédictions imprécises. Assurons-nous donc de fournir à l'IA une alimentation adéquate pour optimiser sa performance.

L'IA utilise plusieurs sources de données pour fonctionner comme les données captées en temps réel grâce à des satellites ou des appareils connectés. Cependant, pour développer un modèle d'IA, les données historiques sont souvent la principale source utilisée. Ces données historiques sont des informations provenant de cas passés, reflétant des événements, des décisions ou des comportements antérieurs. Ces données permettent à l'algorithme de repérer des tendances et des régularités, afin de faire des prédictions ou de générer des résultats en fonction des objectifs définis lors de la première étape du pipeline de l'IA.

Une grande partie de la recherche en apprentissage automatique se concentre sur la création d'algorithmes capables de mieux analyser et comprendre des données pour produire des modèles utiles. Ces algorithmes d'apprentissage ont pour objectif principal de créer des modèles qui reflètent, le plus fidèlement possible, les caractéristiques statistiques des données d'entrée, souvent issues de cas historiques. Ensuite, les modèles créés peuvent être ajustés en y ajoutant d'autres informations, comme des connaissances générales sur le sujet,

des détails spécifiques à l'objectif visé ou au contexte dans lequel le modèle sera utilisé (Fazelpour et Danks, 2021). Il est important de noter que, malgré les modifications ou les ajouts apportés au modèle pour le rendre plus juste, moins biaisé et plus adapté à l'objectif, il demeure un risque que les biais présents dans les données historiques soient reflétés dans l'algorithme ou le modèle.

Les biais dans les données peuvent être classés en deux catégories : ceux qui sont liés à la réalité du monde tel qu'il est et ceux qui découlent des méthodes employées pour collecter les données (Danks et London, 2017 ; Mitchell et collab., 2021).

Nous avons déjà abordé les problèmes liés au monde dans lequel nous vivons, lorsque nous avons évoqué les biais sociaux présents dans la réalité actuelle (écart 1 de la figure 1). Souvent appelé **biais historique**, ce type de biais reflète une inadéquation entre la réalité « telle qu'elle est » et les valeurs communes (idéaux normatifs) sur la façon dont le monde « devrait être ». Ainsi, les modèles qui reproduisent les décisions basées sur la réalité passée ou présente, bien que celle-ci soit exacte dans leurs prédictions à partir des données disponibles, restent biaisés par rapport aux valeurs communes actuelles qui peuvent avoir évolué ou que l'on désire voir changer (Baker et Hawn, 2022). Ce type de biais domine les débats récents sur les biais algorithmiques et est souvent résumé par le slogan « biais à l'entrée, biais à la sortie » (*bias in, bias out*) (Mayson, 2019 ; Rambachan et Roth, 2020).

En ce qui concerne les **biais liés aux méthodes de collecte de données**, ils apparaissent lorsque les variables choisies ne représentent pas bien ce qu'elles sont censées mesurer, ce qui peut affecter la qualité des modèles et des

algorithmes développés ensuite. Ces biais ressemblent au biais historique, mais ils s'en distinguent en donnant une image inexacte du monde réel, actuel ou passé, plutôt qu'une image simplement dépassée. Les biais liés aux méthodes de collecte de données peuvent survenir lorsque les variables choisies pour développer un modèle ou un algorithme ne reflètent pas correctement la réalité ou si certaines variables importantes ne sont pas disponibles (écart 2 dans la figure 1). Cela peut conduire à des prédictions incorrectes ou inéquitables, favorisant certains groupes et en désavantageant d'autres, entraînant des conséquences potentiellement injustes ou discriminatoires pour certaines personnes ou communautés (Baker et Hawn, 2022).

Le biais le plus simple de ce type se produit lorsque les données d'entrée disponibles ne reflètent pas fidèlement le monde réel dans toute sa diversité ou représentativité. La représentativité dont il est question ici est importante dans le secteur public. Lorsqu'une administration consulte un échantillon de personnes pour un projet de politique publique, il est essentiel que cet échantillon reflète bien la diversité de la population (en matière d'âge, de genre, de revenus, de localisation, etc.). Par exemple, si une ville souhaite connaître l'opinion des habitants sur un nouveau projet de transport, il lui faudra consulter un groupe diversifié comprenant des personnes de différents quartiers, âges, niveaux de revenus et modes de déplacement. Si seules les personnes vivant en centre-ville sont interrogées, les résultats ne seront pas représentatifs de l'ensemble des habitants et la politique publique pourrait manquer sa cible. Il en est de même avec les données utilisées pour développer un modèle algorithmique.

En recherche, les données sont souvent comparées à une photographie qui capture une partie d'une situation,

sans refléter toute la complexité du monde réel. Ainsi, les données d'entrée utilisées pour développer un algorithme contiennent certaines caractéristiques de la réalité, mais en laissent d'autres de côté, ce qui diminue la précision de leur représentation (Cossette-Lefebvre et Maclure, 2022). Ce biais, souvent appelé **biais d'échantillonnage** ou biais de représentation, conduit l'algorithme à être moins performant pour un groupe sous-représenté et peut entraîner une mauvaise généralisation des résultats (Baker et Hawn, 2022 ; Srinivasan et Chander, 2021 ; Besse et collab., 2021). Par exemple, les amateurs de fraises associent généralement la couleur rouge à la maturité et à la qualité de ce fruit. Cependant, il existe une variété de fraises blanches dont la saveur rappelle celle de l'ananas. En ne connaissant pas l'existence de cette variété, les amateurs de fraises pourraient laisser ces fraises dans leur assiette, croyant à tort qu'elles ne sont pas mûres[2]. Cette erreur montre comment une connaissance limitée de la réalité peut conduire à des jugements incorrects.

Un autre type de biais associé aux procédures de collecte des données est le **biais de mesure**. Dans certains cas, il est difficile et coûteux, voire impossible, de recueillir des données exactes sur ce que l'on veut mesurer. Pour surmonter ce problème, on va utiliser des données ou des indicateurs intermédiaires (*proxy*). Par exemple, si on ne peut pas mesurer directement le niveau de corruption dans un pays, on peut mesurer, à l'aide d'un sondage, la perception de la population à ce sujet. Il s'agit là d'un moyen indirect d'estimer le niveau de corruption dans ce pays. De la même façon, les taux d'arrestation sont souvent considérés comme un synonyme

2. Nous remercions Daniel Ezra d'avoir porté cet enjeu à notre attention lors d'un séminaire sur l'éthique de l'IA organisé par l'Université de Séoul en mai 2024.

du taux de criminalité. En se basant sur ces données intermédiaires et ces estimations indirectes, souvent liées à des perceptions subjectives, les algorithmes d'apprentissage automatique risquent de perpétuer et de reproduire les inégalités sociales et politiques existantes (Cossette-Lefebvre et Maclure, 2022 ; Srinivasan et Chander, 2021)[3].

Le biais de mesure peut également provenir d'erreurs humaines ou des habitudes des personnes chargées de la saisie des données. Par exemple, lors de la création d'ensembles de données d'images ou de vidéos, les techniques des photographes peuvent limiter la perspective, si les objets sont photographiés sous certains angles seulement. De plus, le biais de mesure peut aussi être causé par l'équipement utilisé pour collecter les données ; une caméra défectueuse pourra produire des images de mauvaise qualité et ainsi fausser les résultats (Srinivasan et Chander, 2021).

Pour être en mesure de fonctionner correctement, une IA doit passer par un processus d'apprentissage, et l'étiquetage est une étape importante dans ce processus. L'étiquetage consiste à associer des descriptions (appelées « étiquettes ») à des données brutes, comme des images ou des textes. Cela permet à l'IA de comprendre ce qu'elle doit reconnaître. Par exemple, pour entraîner une IA à reconnaître des chats dans

3. « Dans les systèmes d'apprentissage automatique adaptatifs, l'algorithme peut être bien programmé, mais il fait ses propres inférences lors de son "apprentissage" pour produire des résultats comme la solvabilité, le risque de maladie cardiaque basé sur des prédispositions génétiques ou des besoins en matière de soins de santé. […] Par exemple, une application d'IA pourrait utiliser les codes postaux pour estimer la classe socioéconomique et prédire le risque de cardiomyopathie de patients. Cette application pourrait alors donner le même score à deux patients du même quartier, bien qu'ils puissent avoir des besoins de soins et des facteurs de risque très différents. Ce type de biais algorithmique peut entraîner une discrimination dans le diagnostic et le traitement des patients marginalisés » (Henderson et collab., 2022, p. 481, trad.).

des images, on montre à l'IA des photos et on leur associe l'étiquette « chat ». C'est donc un peu comme pour apprendre à un enfant à nommer ce qu'il voit en lui montrant des exemples et en lui disant ce que c'est. Grâce à ces exemples, l'IA apprend à reconnaître les chats dans de nouvelles images qu'elle n'a jamais vues. Cette étape est souvent réalisée par des humains qui identifient et nomment les éléments importants dans les données. L'étape de l'étiquetage est également susceptible d'introduire des biais dans les modèles ou les algorithmes (Corbett-Davies et collab., 2023). Ainsi, le **biais d'étiquetage** résulte d'incohérences dans le processus d'annotation. Les styles et préférences des annotateurs peuvent différer, ce qui se traduit par des étiquettes différentes pour la même chose (par exemple, herbe ou pelouse, peinture ou image) (Srinivasan et Chander, 2021).

Un autre type de biais d'étiquetage survient lorsque les préjugés subjectifs des annotateurs influencent le processus. Par exemple, dans une tâche où il s'agit d'annoter les émotions ressenties dans un texte, les étiquettes peuvent être déterminées par les préférences personnelles ou les traits individuels des annotateurs liés à leur culture, leurs croyances ou leur capacité d'introspection. De même, un modèle prédisant la violence à l'école peut être biaisé si les étiquettes des élèves violents sont créées dans un processus empreint de préjugés. Par exemple, imaginons deux événements identiques dans une cour d'école. Pour un élève appartenant à une communauté perçue comme plus « à risque », l'annotateur décrit la situation comme une scène de violence et, pour un autre élève, issu d'une communauté perçue de manière plus favorable, l'annotateur n'associe pas la situation à de la violence. Le même comportement violent de deux élèves est considéré comme de la violence pour les membres d'une communauté, mais pas pour les membres

d'une autre communauté. À terme, cela fausse le modèle en créant un apprentissage basé sur des préjugés des annotateurs, ce qui peut conduire à des prédictions injustes, comme cibler certains groupes d'élèves plus que d'autres sans justification objective. Dans la situation que nous venons de décrire, l'erreur dans l'annotation résulte du **biais de confirmation**, qui est la tendance humaine à rechercher, interpréter, privilégier et retenir les informations confortant ses idées préconçues. Cela signifie que les annotateurs peuvent classer des comportements ou des données selon leurs propres perceptions empreintes d'idées reçues, de croyances et de préjugés, plutôt que de s'appuyer sur une analyse impartiale (Lloyd et Hamilton, 2018 ; Srinivasan et Chander, 2021).

Les erreurs d'étiquetage peuvent également être causées par l'**effet de pointe** ou l'effet de récence, un biais cognitif lié à la mémoire, qui amène les individus à évaluer une expérience principalement en fonction de son moment le plus intense et de sa conclusion, plutôt que sur l'ensemble ou la moyenne des événements vécus. Par exemple, certains annotateurs peuvent accorder davantage d'importance à la dernière partie d'une conversation lorsqu'ils choisissent une étiquette afin de qualifier une conversation (Srinivasan et Chander, 2021).

Les biais d'étiquetage peuvent également être causés par les « micro-travailleurs » ou « travailleurs de plateformes » (*gig workers*) qui participent à leur développement. Ces travailleurs accomplissent diverses tâches, telles que l'étiquetage de données, l'annotation d'images ou encore la validation des résultats que nous aborderons plus loin. Parce qu'ils sont souvent aux prises avec des conditions précaires et un rythme de travail infernal, leurs annotations sont produites d'une manière expéditive et parfois influencées par leurs

biais cognitifs ou culturels (Wang et collab., 2022). Ces biais humains se transmettent alors aux systèmes d'IA, qui les reproduisent ou les amplifient. La pénibilité de ces conditions de travail a été mise en lumière, en particulier en raison de leur exposition à des images ou des scènes traumatisantes et d'une violence extrême dans le cadre de leur travail d'étiquetage des données (Williams et collab., 2022).

Finalement, un autre élément peut engendrer des biais dans les données. Il s'agit du **biais d'étiquetage sélectif**, qui fait référence à une situation où un modèle est entraîné à partir d'un ensemble de données qui ne contient que certaines étiquettes ou classifications. Par exemple, si un modèle de reconnaissance faciale est formé principalement sur des visages d'une seule origine ethnique, il aura du mal à bien reconnaître des visages d'autres origines. L'étiquetage sélectif des données d'entraînement crée un biais dans les résultats (Fazelpour et Danks, 2021).

Un phénomène similaire se produit avec le **biais d'ensemble de données négatif**, lorsque les données d'entraînement manquent d'exemples d'instances négatives, c'est-à-dire des situations où le phénomène n'est pas observé. Par exemple, si un modèle d'IA pour évaluer le risque d'insolvabilité dans un programme de microcrédit est formé uniquement avec des données de personnes ayant obtenu un prêt et l'ayant remboursé, sans inclure suffisamment d'exemples de personnes n'ayant pas remboursé le prêt, le modèle ne pourra pas correctement faire ressortir les caractéristiques des emprunteurs à risque. Cela entraînera des prédictions biaisées et une évaluation inadéquate du risque d'insolvabilité. Une diversité d'étiquettes, incluant des instances à la fois positives et négatives, est essentielle pour parvenir à formuler des prédictions fiables (Torralba et Efros, 2011).

2.3 Biais dans la modélisation et la validation

La modélisation d'une IA consiste à développer un algorithme adapté à la résolution du problème défini. La modélisation débute par une phase d'entraînement durant laquelle l'algorithme utilise les données disponibles pour apprendre à faire des prédictions ou prendre des décisions. Ce jeu de données initial est appelé ensemble d'apprentissage ou d'entraînement (*training set*). Pendant cet entraînement, les développeurs définissent et ajustent les paramètres du modèle afin d'améliorer ses performances. Une phase de validation suit, durant laquelle l'IA est testée sur un autre ensemble de données inédites pour évaluer sa capacité à faire des prédictions précises dans des situations nouvelles. Cet ensemble de données, appelé ensemble de test (*test set*), n'a jamais été traité par le modèle auparavant. La validation permet de garantir que les résultats obtenus correspondent aux attentes. La vérification a pour objectif de s'assurer que l'algorithme ne produit pas de résultats erronés, incomplets ou imprévus. Si les résultats des tests sont satisfaisants, l'IA est alors prête à être utilisée dans des applications concrètes.

Ces processus de modélisation et de validation sont souvent réalisés de manière itérative par les développeurs afin d'évaluer et d'améliorer les performances du modèle en fonction de critères de réussite définis par les développeurs ou par les commanditaires de l'algorithme (Fazelpour et Danks, 2021).

Même si cette étape de modélisation et de validation peut être très technique, elle implique tout de même de nombreux jugements de valeur, ce qui risque d'introduire des biais.

Parmi les biais susceptibles de survenir lors des phases de modélisation et de validation figurent notamment les **biais liés aux ensembles de données** d'entraînement et de test.

Les biais déjà repérés dans les données à l'étape précédente peuvent également se manifester ici et avoir des répercussions majeures sur les résultats finaux de l'algorithme. Tout comme notre alimentation influence notre performance physique – il est bien plus difficile de courir un marathon après avoir englouti une poutine et une boisson gazeuse –, les données biaisées affectent directement la qualité des résultats produits par l'algorithme.

Un des éléments qui peut affecter la qualité des ensembles de données utilisés pour développer le modèle est le **biais de traitement de l'échantillon**. Ce biais survient lorsque l'ensemble de données utilisées pour concevoir ou évaluer l'algorithme n'est pas représentatif de la population cible. Cela peut se produire si les données d'entraînement ou de test sont biaisées ou contiennent des caractéristiques particulières qui ne reflètent pas celles de la population où sera utilisé l'algorithme. Ce manque de représentativité peut fausser les résultats et réduire la performance de l'algorithme lors de son déploiement à grande échelle. Par exemple, dans le cas d'un algorithme conçu pour proposer des services à des usagers, certains d'entre eux peuvent recevoir une information au sujet d'un service ou d'un programme, tandis que d'autres n'y auront pas accès. Si les tests de validation sont effectués principalement avec des données d'utilisateurs bilingues par exemple, cela peut influencer le comportement final de l'algorithme. En effet, lorsqu'il sera déployé à grande échelle auprès d'une population majoritairement unilingue, les recommandations ou les résultats produits pourraient être biaisés, ne reflétant pas la composition linguistique de la population. Ce biais est comparable au biais d'échantillonnage mentionné à l'étape de la collecte des données. Le biais de traitement de l'échantillon survient cependant au moment de l'évaluation du fonctionnement d'un modèle

algorithmique, et non à l'étape de la collecte initiale des données (Srinivasan et Chander, 2021).

Lors de la modélisation, des biais peuvent aussi résulter de limitations techniques propres à l'algorithme ou de contraintes internes au système, telles que la capacité de calcul. Par exemple, si des logiciels censés distribuer des résultats de manière aléatoire privilégient des éléments en début ou en fin de liste, ils risquent de biaiser le modèle (Friedman et Nissenbaum, 1996 ; Srinivasan et Chander, 2021).

Un autre type de biais qui affecte la qualité de la performance de l'algorithme est le **biais d'évaluation de la performance**. Ce biais se produit lorsque les méthodes ou les mesures utilisées pour juger la performance d'un algorithme ne sont pas adaptées aux objectifs du modèle. Cela arrive lorsque les données ou les critères choisis pour évaluer le modèle ne permettent pas de vérifier correctement son efficacité dans le monde réel (Srinivasan et Chander, 2021). Par exemple, un algorithme peut sembler performant en obtenant de bons résultats sur une mesure générale, comme la précision globale. Cela ne garantit cependant pas que le modèle sera performant sur l'ensemble des données testées. Par exemple, si un modèle fait 100 prédictions et que 90 sont justes, sa précision globale est de 90 %. Cependant, cette mesure pourrait être trompeuse si certaines catégories de la population sont sous-représentées dans les données d'entraînement. Par conséquent, l'algorithme pourrait générer de moins bons résultats pour ces groupes minoritaires, même si ses performances globales sont bonnes. De nombreux modèles utilisés pour analyser des données en éducation sont souvent basés sur des échantillons qui ne reflètent pas la diversité réelle des élèves (Paquette et collab., 2020). De plus, de nombreux articles scientifiques qui présentent des modèles algorithmiques ne précisent pas

sur quelles populations les modèles ont été testés. Cela rend difficile, voire impossible, de détecter des biais d'évaluation, car, sans ces informations, on ne sait pas si le modèle a été évalué sur un groupe représentatif de la population ou non (Baker et Hawn, 2022). Par exemple, un modèle d'analyse de la réussite scolaire pourrait sembler efficace, mais, si on ne sait pas s'il a été testé sur une population diversifiée (incluant différentes origines sociales ou géographiques), il n'est pas possible de vérifier s'il fonctionne correctement pour tous les groupes. Sans cette information, les biais qui affectent certaines minorités pourraient passer inaperçus.

Les chercheurs en administration publique savent que le choix des critères et des indicateurs pour évaluer la performance des interventions publiques n'est pas un exercice neutre et que certains critères retenus peuvent se contredire. Il en est de même dans le domaine de l'IA. Lors de la modélisation et de la validation d'un algorithme, il est parfois impossible d'optimiser simultanément plusieurs paramètres, car certains critères de performance entrent en conflit les uns avec les autres. Cela oblige les concepteurs à faire des compromis pour surmonter le **problème d'optimisation multi-objectif** (Kleinberg, Lakkaraju et collab., 2017 ; Friedler et collab., 2016). En déterminant quel critère sera priorisé, on introduit un jugement de valeur dans le processus de validation de l'algorithme, influençant ainsi le choix des mesures statistiques qui seront ou non privilégiées (Fazelpour et Danks, 2021). Par exemple, dans le développement d'un algorithme visant à contrer la fraude fiscale, il existe une tension entre la fiabilité et l'adaptabilité. L'entraînement du modèle sur des données historiques vise à maximiser sa fiabilité en s'appuyant sur les schémas de fraude connus. Toutefois, cela limite son efficacité face aux nouvelles formes de fraude. Pour accroître l'adaptabilité, il serait possible

d'intégrer des données non validées ou des mécanismes d'apprentissage continu, mais cela pourrait réduire la fiabilité des résultats produits par l'algorithme. Les choix qui sont faits à cette étape ne sont donc pas purement techniques, mais reflètent les objectifs et les valeurs des concepteurs et des utilisateurs du modèle (Kearns et Roth, 2020).

Dans le système Robodebt, conçu pour détecter les trop-perçus dans les prestations sociales en Australie, les algorithmes ont été mis en place sans vérifier si leurs calculs pouvaient entraîner une surévaluation des montants dus. La précision a été compromise au profit de la rapidité (Rinta-Kahila, 2024).

Dans le cas du système COMPAS, qui estimait la probabilité de récidive des détenus aux États-Unis, il a fallu faire un choix entre la parité de la valeur prédictive et l'égalité des taux d'erreur. La parité de la valeur prédictive signifie que, lorsque l'algorithme prédit un résultat, la répartition des prédictions correctes doit être proportionnellement équivalente pour tous les sous-groupes considérés. Cela permet d'éviter que le programme soit plus précis dans ses prédictions pour certains groupes et moins fiable pour d'autres. L'égalité des taux d'erreur, quant à elle, signifie que, lorsque l'algorithme se trompe, il doit faire des erreurs de manière équivalente pour tous les groupes (en nombre de faux positifs et de faux négatifs). Cela garantit que l'algorithme ne fait pas plus d'erreurs pour un groupe que pour un autre.

Pour illustrer la difficulté rencontrée par les développeurs face au problème d'optimisation multi-objectif, supposons que COMPAS évalue deux types de détenus : des délinquants sexuels et des cambrioleurs. La probabilité qu'une personne récidive est généralement plus basse chez les délinquants sexuels que chez les cambrioleurs, ce qui signifie qu'ils ont

des taux de base différents. Si l'algorithme est conçu pour être plus précis dans ses prédictions pour les cambrioleurs, il pourrait prédire la récidive avec plus de fiabilité pour eux que pour les délinquants sexuels. Dans ce cas, on ne respecterait pas le critère de la parité de la valeur prédictive, car la précision varie entre les deux groupes de détenus. À l'inverse, si l'algorithme se trompe plus souvent en classant à tort les délinquants sexuels comme « à risque de récidive », mais se trompe moins fréquemment pour les cambrioleurs, cela signifie que les taux d'erreur ne sont pas égaux entre les groupes. Dans la pratique, il est très difficile d'ajuster un algorithme pour qu'il optimise simultanément la précision des prédictions pour tous les groupes, tout en garantissant que les erreurs soient réparties de manière égale. Ce choix est encore plus difficile lorsque, dans une population de détenus, le taux de récidive varie entre des groupes de délinquants condamnés pour des crimes différents. Cette différence entre les groupes rend difficile l'équilibrage des critères de parité prédictive et d'égalité des taux d'erreur, car les algorithmes auront tendance à se comporter différemment selon les groupes en fonction des écarts initiaux. Les concepteurs doivent donc choisir quelle valeur privilégier, ce qui conduit à procéder à un arbitrage entre l'équité des prédictions et l'équité des erreurs. Cette décision risque de reproduire des inégalités et d'accroître la discrimination.

Les personnes chargées d'évaluer les performances des modèles d'IA peuvent elles aussi introduire des **biais humains de l'évaluation**, en raison de leurs propres préjugés, de biais cognitifs ou des conditions de travail dans lesquelles elles se trouvent. Comme pour les biais liés à la collecte des données que nous avons présentés précédemment, des phénomènes comme le biais de confirmation (tendance à privilégier les informations qui confirment des croyances

préexistantes) et l'effet de récence (accorder plus de poids aux dernières informations obtenues) peuvent fausser l'évaluation. De plus, les évaluateurs sont limités par leur mémoire, ce qui peut entraîner un **biais de mémorisation**, en raison du temps réduit dont ils disposent pour accomplir leur travail, ainsi que par la fatigue cognitive due aux exigences de leurs conditions de travail (Srinivasan et Chander, 2021).

Finalement, mentionnons que, lors de la modélisation et de la validation d'un modèle algorithmique, des efforts peuvent être faits pour corriger des biais historiques ou sociaux qui existent entre le monde idéal et le monde réel, représentés par l'écart 1 de la figure 1. Cependant, ces efforts pour atténuer ou supprimer ces biais sociaux peuvent à leur tour introduire de nouveaux biais. Pour comprendre ce paradoxe, il faut se rappeler que les initiatives visant à accroître l'équité, qui sont de plus en plus valorisées au sein de la société en général et dans l'écosystème de l'IA en particulier, introduisent des critères de performance supplémentaires à l'étape de la modélisation et de la validation des algorithmes. L'équité peut être définie et mesurée de différentes manières selon la théorie normative retenue (par exemple, égalité des résultats ou équité procédurale). Ces variations conceptuelles influencent la manière dont ce critère sera intégré dans les modèles d'IA (Corbett-Davies et collab., 2017 ; Kleinberg, Ludwig et collab., 2018). C'est ainsi que la quête de l'équité, qui s'appuie sur des algorithmes, impose souvent des compromis privilégiant certaines valeurs par rapport à d'autres. Même si les outils mathématiques permettent de formaliser ces compromis, leur résolution implique des réflexions philosophiques au-delà des approches techniques de l'équité (*fair ML*). Par exemple, dans le domaine de la santé, un algorithme pourrait être conçu pour répartir le personnel médical selon la gravité de l'état de santé des patients. Pour

garantir l'équité, d'autres critères, comme la priorisation des groupes historiquement marginalisés, peuvent être ajoutés. Cet ajout peut affecter la précision des prédictions et aboutir à des décisions qui favorisent les patients de ces groupes marginalisés. Par exemple, un patient moins malade issu d'un groupe défavorisé pourrait recevoir des soins en priorité. Ce compromis entre équité et efficacité soulève des enjeux philosophiques sur la justice sociale et le bien-être collectif. L'atteinte de ce compromis nécessite de concilier de multiples objectifs qui ne sont pas toujours facilement réconciliables. En transposant le débat dans le domaine de l'IA, nous comprenons facilement que l'enjeu devient encore plus complexe, surtout si ces aspects ont été négligés par les responsables politiques et les gestionnaires publics lors de la première étape du pipeline de l'IA, à savoir la définition du problème. Ainsi, la conception des systèmes d'IA devrait être le fruit d'une réflexion approfondie et englobante sur les principes de bien-être et d'équité. Cependant l'intégration de cette perspective holistique dans la conception des algorithmes représente un défi pour les développeurs, qui ont tendance à se concentrer sur une partie du problème et cherchent à modéliser la solution d'une manière essentiellement technique et mathématique.

Une fois que le modèle d'IA a été ajusté et testé pour vérifier s'il fonctionne bien et donne des résultats satisfaisants, il est prêt à être déployé et utilisé dans des situations réelles.

2.4 Biais dans le déploiement

La dernière étape du pipeline de l'IA est le déploiement, qui correspond à sa mise en fonction et à son utilisation dans le monde réel. Dans le secteur public, cela signifie que l'IA est intégrée au sein de l'administration pour traiter des

dossiers, comme l'attribution d'aides sociales ou la gestion des demandes de permis. Afin de garantir que le système d'IA fonctionne adéquatement, il est nécessaire de surveiller ses performances en situations réelles, d'ajuster ses paramètres et de mettre à jour ses données régulièrement.

C'est principalement au moment du déploiement d'un système algorithmique que les utilisateurs finaux peuvent observer et utiliser les résultats produits. Pour les employés du secteur public ou les responsables politiques, cela signifie que les prédictions ou autres informations fournies par le système d'IA soutiennent leur travail et alimentent ou influencent leurs décisions (De-Arteaga, 2020).

À cette étape, l'interprétation et l'utilisation des résultats algorithmiques peuvent varier en fonction de plusieurs facteurs, tels que le contexte décisionnel (Green et Chen, 2019 ; Kleinberg, Lakkaraju et collab., 2017), le degré de confiance dans les algorithmes (Dietvorst et collab., 2015) et les cadres de responsabilité institutionnelle. Par conséquent, des décisions moralement discutables ou injustes ainsi que des préjudices peuvent être causés aussi bien par un algorithme biaisé soutenu par un humain impartial que par un algorithme impartial soutenant un humain biaisé ou résulter de la combinaison des biais humains et algorithmiques.

Comme nous le verrons avec les différents types de biais présentés dans cette section, l'interaction entre l'humain et la machine est également susceptible d'engendrer plusieurs problèmes.

Tout d'abord, le **biais de déploiement** survient lorsqu'un modèle est appliqué à un usage pour lequel il n'a pas été conçu. Par exemple, un modèle développé pour détecter le désengagement des étudiants pendant une formation serait mal utilisé s'il servait à attribuer des notes de participation

à la fin de cette formation (Baker et Hawn, 2022). Ce biais peut donc se produire lorsque le rôle ou le fonctionnement du modèle n'est pas bien compris par ses utilisateurs. Pour illustrer cet élément, prenons l'exemple d'un modèle purement prédictif qui serait utilisé pour modifier des politiques ou prendre des décisions. Si un algorithme prédit qu'un étudiant risque d'obtenir de mauvaises notes en se basant uniquement sur ses retards fréquents, cela ne signifie pas que les retards sont la véritable cause de ses résultats scolaires. Il pourrait y avoir des facteurs sous-jacents, comme des difficultés familiales ou un manque de concentration en classe, qui influencent à la fois ses retards et ses notes. S'appuyer exclusivement sur la variable du retard pour justifier les décisions pourrait alors mener à des interventions mal ciblées auprès des étudiants en retard.

Ce raisonnement s'applique également aux algorithmes qui prédisent le taux de criminalité en se basant sur des variables, comme le niveau de pauvreté ou le taux de chômage de certains quartiers. Ces variables ne sont pas nécessairement la cause directe de la criminalité. Si les décideurs publics interprètent cette prédiction comme un lien de cause à effet, ils risquent de concentrer leurs actions sur ces éléments en pensant que cela suffira à résoudre le problème. Cependant, la réalité est souvent bien plus complexe. D'autres facteurs, comme l'accès à l'éducation, le manque de services publics ou la qualité des infrastructures, peuvent jouer un rôle sur l'augmentation de la criminalité. En focalisant uniquement sur les variables identifiées par l'algorithme, les décideurs pourraient négliger ces autres aspects et mettre en place des politiques inefficaces, voire contre-productives. L'algorithme offre donc une prédiction basée sur une corrélation, mais cette corrélation ne doit pas être confondue avec une relation de cause à effet. Une approche holistique est

nécessaire pour comprendre les véritables causes du phénomène et pour élaborer des politiques publiques pertinentes et efficaces. Dans cet exemple, une mauvaise compréhension du fonctionnement de l'algorithme par les responsables politiques pourrait en plus aggraver la stigmatisation dont sont déjà victimes les habitants des quartiers défavorisés. Il est donc essentiel que les prédictions générées par l'algorithme soient communiquées de façon claire aux utilisateurs afin qu'ils puissent les utiliser correctement en tenant compte de leurs limites et de leurs véritables capacités, qu'elles soient prédictives, diagnostiques ou causales (De-Arteaga, 2020).

Le biais de déploiement peut aussi s'expliquer par le fait que plusieurs utilisateurs orientent leurs comportements ou leurs décisions en fonction des informations fournies par un algorithme, sans bien comprendre les raisons pour lesquelles l'algorithme a été conçu ou en interprétant de manière inadéquate les résultats produits. Par exemple, la Société des alcools du Québec (SAQ) pourrait utiliser un algorithme pour prédire le type de vin que sa clientèle est la plus susceptible d'acheter, afin de mieux planifier son approvisionnement. Si un client interprète ces résultats comme des recommandations personnalisées sur les vins qu'il pourrait apprécier, il risque d'être déçu par les suggestions. Avec cet exemple, nous voyons que les mêmes résultats produits par un algorithme peuvent être pertinents pour la SAQ, mais trompeurs pour sa clientèle. Les erreurs et biais algorithmiques proviennent donc, dans ce cas, de la manière dont les utilisateurs interprètent les résultats. Un client comprendrait mieux les suggestions algorithmiques s'il réalisait que l'algorithme reflète les tendances générales d'achat de la clientèle de la SAQ, plutôt que ses préférences individuelles. Le biais de déploiement découle ainsi d'une incompréhension des

objectifs initiaux de l'algorithme et de la manière correcte d'interpréter ses résultats (Sénéchal, 2023).

Dans le même ordre d'idées, nous constatons que les algorithmes optimisent les valeurs sur lesquelles ils ont été formés. Ainsi, des biais importants peuvent émerger lorsque les valeurs des utilisateurs divergent de celles de l'algorithme. Ce biais, défini comme le **biais d'alignement des valeurs**, se produit lorsque l'algorithme ne prend pas correctement en compte les valeurs et les attentes des utilisateurs ou de la société (Danks et London, 2017). Prenons l'exemple d'un programme gouvernemental de soutien aux personnes en recherche d'emploi qui utilise un algorithme pour prédire quelles personnes sont susceptibles de retrouver un emploi rapidement. L'algorithme se base sur des critères comme le niveau d'éducation et l'expérience professionnelle pour maximiser le taux de retour à l'emploi. Cependant, l'algorithme pourrait négliger des facteurs importants pour les responsables politiques, comme la qualité de l'emploi trouvé ou les besoins particuliers des personnes les plus vulnérables, par exemple celles qui sont en situation de handicap ou qui vivent avec un problème de santé mentale. En optimisant uniquement le retour rapide à l'emploi, l'algorithme risque d'orienter les ressources vers les personnes qui ont déjà de meilleures chances de s'en sortir et de laisser de côté les personnes en recherche d'emploi qui doivent surmonter des obstacles plus importants. Ainsi, le biais d'alignement des valeurs peut entraîner une prise de décision qui, bien qu'elle soit techniquement conforme aux prévisions de l'algorithme, manque de pertinence sociale et aggrave les inégalités.

Le **biais de classement** est étroitement lié à la question de l'alignement des valeurs. L'ordre de présentation des informations générées par un algorithme reflète les priorités et les valeurs implicites de ses concepteurs. Cet ordre

influence ensuite la perception et la décision de l'utilisateur. Par exemple, Google classe les résultats de recherche en fonction de critères de pertinence définis par son algorithme. Les utilisateurs ont tendance à cliquer sur les premiers liens sans parcourir l'ensemble des liens proposés. Ce phénomène met en évidence l'alignement implicite des valeurs qui tend à favoriser certains contenus mis en avant par Google (parfois pour lui garantir davantage de revenus publicitaires) au détriment des besoins ou des intérêts spécifiques des utilisateurs (Srinivasan et Chander, 2021). Ce biais peut se manifester de différentes façons dans le secteur public. Prenons l'exemple d'un site gouvernemental qui permet aux parents de rechercher des informations sur les écoles publiques de leur région. L'algorithme de recherche pourrait classer les résultats en fonction de critères tels que les résultats aux examens du ministère, le taux de diplomation ou le ratio élèves-enseignants. Si l'algorithme accorde une grande importance aux résultats aux examens, il privilégie implicitement la valeur de la performance scolaire. Par conséquent, les écoles en tête de liste seront probablement celles qui ont les meilleurs scores aux examens, sans tenir compte d'autres aspects qui pourraient être importants aux yeux des parents, comme la présence de programmes artistiques ou sportifs, la proximité du domicile ou le soutien offert aux élèves à besoins particuliers, par exemple.

De plus, lors de l'intégration d'un système d'IA au sein d'une organisation publique, il est important de s'assurer que l'environnement dans lequel un algorithme est déployé correspond au contexte dans lequel les données d'entraînement ont été collectées ou, au moins, de comprendre les différences entre ces contextes. Lorsqu'un modèle, formé dans un environnement spécifique, est déployé dans un autre contexte sans ajustements appropriés, cela peut entraîner

des **biais de transfert**. Ce type de biais survient lorsque les données utilisées pour entraîner un modèle diffèrent de celles qu'il utilise lors de son déploiement dans le monde réel. Cela peut entraîner des résultats erronés, car le modèle a appris à partir de données spécifiques, mais les conditions réelles dans lesquelles il est implanté sont différentes. Des chercheurs observent ces biais dans les établissements d'enseignement supérieur qui utilisent des algorithmes pour prédire la réussite étudiante. Pour cela, de plus en plus d'universités utilisent les données d'anciens étudiants pour construire des modèles prédictifs de la réussite de la population étudiante actuelle ou future. Selon les modèles, la définition de la réussite et les critères utilisés pour la quantifier pourront varier. Malgré ces différences, de nombreuses administrations universitaires utilisent les résultats générés par ces modèles pour alimenter leur réflexion et appuyer des réformes de leurs politiques d'admission ou de leurs programmes de soutien à la réussite. Cependant, un algorithme prédictif de réussite étudiante conçu à partir de données historiques d'une petite université privée des États-Unis ne pourra pas nécessairement être intégré dans une grande université publique en France (Fazelpour et Danks, 2021). Cet exemple souligne l'importance de connaître le contexte dans lequel un algorithme a été développé et d'en tenir compte afin d'évaluer la pertinence de l'utiliser dans un pays ou au sein d'une organisation publique particulière.

Le **biais d'interaction** peut quant à lui survenir lorsque des systèmes algorithmiques sont conçus pour apprendre sans qu'aucune mesure de protection soit mise en place pour détecter et exclure les comportements inappropriés des utilisateurs. Ces derniers peuvent alors détourner complètement les finalités initiales du système. L'exemple le plus emblématique est celui du Chatbot Tay de Microsoft, qui

avait pour objectif de dialoguer de manière informelle sur les réseaux sociaux et de démontrer les capacités de l'intelligence artificielle à s'adapter aux interactions humaines, en fonction des données recueillies à partir des échanges en ligne. L'expérience a été de courte durée : moins de 24 heures après son lancement, le projet a dû être interrompu. En l'absence de garde-fous adéquats, Tay s'est rapidement mis à proférer des insultes racistes et antisémites apprises à partir des interactions avec des utilisateurs (Lloyd et Hamilton, 2018). Par la suite, d'autres systèmes d'IA ont connu le même problème.

Des **biais de rétroaction** peuvent survenir après le déploiement d'un algorithme. Premièrement, lorsqu'un algorithme influence les décisions futures, ces résultats sont parfois réinjectés dans les données d'entraînement, ce qui renforce les biais déjà présents (Levy et collab., 2021). Par exemple, aux États-Unis, les algorithmes de police prédictive utilisent des données d'arrestations qui proviennent souvent de quartiers majoritairement peuplés de personnes de couleur. Cela conduit à une surreprésentation de ces zones dans les patrouilles policières, entraînant plus d'arrestations et exacerbant la surreprésentation de ces populations dans les prisons (Ensign et collab., 2017 ; Lum et Isaac, 2016). À l'inverse, les quartiers moins ciblés par les données historiques ne font pas l'objet de surveillance, ce qui maintient un déséquilibre dans la répartition des patrouilles policières.

De plus, les algorithmes sont souvent intégrés dans des systèmes sociaux complexes où des boucles de rétroaction influencent à leur tour l'évolution des données et l'exactitude des prédictions algorithmiques. Ces dynamiques, issues de l'interaction entre le système algorithmique et son environnement, peuvent modifier les contextes initiaux de manière importante (Chouldechova et Roth, 2020). Par exemple, les décisions actuelles concernant l'allocation des

ressources aux initiatives de soutien aux étudiants auront des répercussions sur la composition de la population étudiante des prochaines années, ce qui influencera ainsi les prochains résultats scolaires et les futurs besoins en ressources. En déployant un algorithme dans un tel contexte, celui-ci peut modifier l'environnement dans lequel il entre en jeu et y introduire des biais. En effet, ces systèmes automatisés ne fonctionnent pas de manière isolée, mais ils s'insèrent dans un écosystème plus vaste. Ces interactions peuvent générer des boucles de rétroaction complexes, où les actions d'un algorithme influencent le comportement des institutions, qui en retour modifient l'environnement de l'algorithme, créant ainsi des effets en cascade où les pratiques humaines et algorithmiques se façonnent et s'influencent mutuellement (Ensign et collab., 2017 ; Milli et collab., 2018).

Cet enchevêtrement peut avoir des conséquences inattendues lorsque les individus, les organisations ou les groupes ciblés modifient leurs comportements de manière délibérée, voire stratégique, en fonction de leur compréhension du fonctionnement de l'algorithme (Dai et collab., 2021). Par exemple, si un établissement d'enseignement supérieur utilisait les résultats d'un algorithme prédictif pour établir la liste des étudiants admissibles à un programme de soutien et de remédiation scolaire, certains étudiants proches du seuil d'admissibilité à ce programme pourraient être tentés d'obtenir délibérément de moins bonnes notes afin de bénéficier du soutien offert. Cette situation illustre un phénomène bien connu des experts en mesure de la performance : les individus soumis à des indicateurs de performance adaptent souvent leur comportement, ou cherchent à manipuler les indicateurs, pour se conformer aux attentes. L'économiste britannique Charles Goodhart a résumé ce risque dans une loi : « lorsqu'une mesure devient un objectif, elle cesse

d'être une bonne mesure ». Cette loi de Goodhart s'applique également aux systèmes algorithmiques et devrait inciter leurs concepteurs à réfléchir aux éventuelles distorsions et conséquences imprévues que les systèmes algorithmiques peuvent engendrer.

Lors de la phase de déploiement d'un système algorithmique, il est donc important de garder à l'esprit que les analyses prédictives ou les autres résultats générés par l'IA perdent de leur utilité lorsque les acteurs les instrumentalisent à leur avantage ou mettent en place des stratégies pour les contourner.

> C'est ainsi que les écoles enseignent en fonction des tests, que les universités manipulent les classements et que les personnes à la recherche d'un emploi apprennent quels signaux inclure dans leurs CV pour déjouer les filtres automatisés. Ces formes de rétroaction peuvent devenir des sources de tension, car les personnes affectées par un système tentent de s'y adapter et d'anticiper ses attentes, tandis que les concepteurs du système essaient de s'assurer [qu'il demeure pertinent et] qu'il fonctionne comme prévu (Levy et collab., 2021, p. 326, trad.).

Ce jeu du chat et de la souris s'observe notamment dans le secteur social, où des intervenants travaillant avec des familles en difficulté ont établi des stratégies pour contourner les recommandations algorithmiques en adaptant les informations qu'ils transmettent à l'algorithme, afin d'obtenir des résultats qui correspondent à leur propre évaluation de la situation. Ce comportement révèle un conflit sous-jacent entre la prise de décision automatisée et l'expertise humaine. Lorsqu'un algorithme entre en contradiction avec le jugement professionnel d'un employé, ce dernier pourrait être tenté de manipuler le système en modifiant les données d'entrée ou en influençant le processus afin d'obtenir une recommandation qui se rapproche davantage de sa propre appréciation de la situation, à la lumière de son jugement

professionnel. Avec ce biais de rétroaction, nous constatons que les intervenants sociaux cherchent à se réapproprier le pouvoir de décision que l'algorithme leur retire en partie (Jacob et Souissi, 2022).

Dans d'autres situations, l'interaction entre l'humain et la machine va parfois aboutir à privilégier l'outil technologique. C'est ce qui se produit avec le **biais d'automatisation**, qui se manifeste par la tendance à se fier aux outils technologiques, malgré des signaux contradictoires. Par exemple, en suivant aveuglément les indications d'un GPS, nous pourrions nous retrouver dans une impasse ou sur un chemin impraticable, même si des indices visuels nous invitent à remettre en question l'itinéraire proposé. Ce biais est à l'œuvre lorsque les décideurs délèguent entièrement ou partiellement leur pouvoir décisionnel à des systèmes automatisés, sans remettre en question ni vérifier les résultats produits. En se reposant sur le pilote automatique, ces décideurs renoncent à leur jugement critique.

En Espagne, les forces de l'ordre ont mis en place le dispositif VioGén, un système de suivi des cas de violence conjugale[4]. Ce dispositif évalue le risque posé par un agresseur en tenant compte de divers facteurs liés à la situation du couple. Comme l'explique la commissaire en chef responsable de VioGén, le système « regarde s'il y a des mineurs à charge, s'il existe des conflits, des dépendances à la drogue ou des problèmes de santé mentale. [Il] regarde également comment se sont produits les faits [rapportés par la victime], quel type d'agression a subi la femme, tout cela est analysé. Ce n'est pas la même chose si la victime a été

4. Nous remercions Réjean Roy, directeur de la formation et de la mobilisation des connaissances chez IVADO d'avoir attiré notre attention sur cet exemple.

attrapée par le cou ou par le bras » (Demon, 2023). Le niveau de risque calculé par l'algorithme détermine les mesures de protection policière qui sont automatiquement activées par VioGén. Cependant, ce système a suscité des controverses, notamment à cause de tragédies qui ont attiré l'attention des médias et du public. Un audit de l'algorithme a révélé que, « même si VioGén est conçu comme un système de recommandation, les policiers s'écartent rarement du score de risque attribué automatiquement, ce qui fait que c'est le système algorithmique qui dicte des mesures de protection pour les victimes dans la majorité des cas » (Eticas.ai, 2023, trad). En effet, cet audit nous apprend que seulement 5 % des scores de risque ont été ajustés à la hausse par des policiers. Cette confiance excessive dans les résultats du système algorithmique a malheureusement coûté la vie à plusieurs femmes.

Le biais d'automatisation se manifeste également chez les juges américains qui se basent sur les scores de risque générés par l'IA pour déterminer la libération d'un détenu. Toutefois, dans cette situation, il est utile de mentionner que l'existence même du score de risque dans le dossier d'une personne incarcérée influence le processus décisionnel du juge :

> En effet, les juges ont des incitations à ne pas ignorer les recommandations automatisées. Imaginez qu'un juge décide de libérer un accusé malgré un score de risque automatisé élevé et que cet accusé commette ensuite un crime après sa libération. Le juge pourrait faire face à des réactions négatives et à des critiques, étant donné qu'il existe désormais dans le dossier un score de prédiction apparemment précis que le juge a choisi de laisser de côté. Le choix le plus sûr pour le juge est simplement d'adopter la recommandation automatisée, car il peut toujours se référer à ce score de risque comme justification de sa décision (Surden, 2020, p. 734, trad.).

Le biais d'automatisation pourrait s'amplifier à mesure que les décideurs intègrent l'IA dans leur quotidien et qu'ils utilisent des systèmes d'IA de plus en plus performants. Cette familiarité croissante pourrait rendre les décideurs de plus en plus dépendants aux algorithmes et les amener à ne plus remettre en question les résultats produits. Les décideurs risquent également de supposer, à tort, que tous les systèmes d'IA offrent le même niveau de fiabilité, sans évaluer les variations des taux de précision entre les systèmes algorithmiques qu'ils utilisent. Cela est préoccupant, car l'idée de maintenir les humains dans la boucle est perçue comme un moyen essentiel de contrôler les erreurs et d'éviter les biais algorithmiques. Les exemples précédents nous enseignent que les décideurs et les agents publics ont tendance à suivre les recommandations de l'algorithme, même lorsqu'ils devraient probablement les remettre en question. Cela révèle une faiblesse dans leur capacité à exercer un contrôle effectif sur les résultats produits par l'algorithme (De-Arteaga et collab., 2020). « Ces préoccupations deviennent particulièrement problématiques dans la prise de décision algorithmique […], lorsque les biais humains rencontrent les biais algorithmiques » (Alon-Barkat et Busuioc, 2023, p. 166, trad.).

Lors de l'évaluation des résultats générés par un algorithme, les utilisateurs devraient connaître son niveau de précision, afin de pouvoir interpréter correctement la fiabilité, l'exactitude et la pertinence des résultats. Tout comme on tient compte de la marge d'erreur pour interpréter la fiabilité d'un sondage (au risque de connaître une grande désillusion lors de la soirée électorale), les utilisateurs d'un algorithme devraient comprendre que les résultats produits ne sont pas infaillibles. Connaître le degré de précision d'un algorithme permet de mieux gérer les attentes et d'éviter de considérer les résultats comme des vérités absolues. De plus,

la réflexion et le discernement humains sont indispensables pour intégrer dans l'analyse des éléments que le système automatisé pourrait omettre. Certains éléments importants d'une situation peuvent en effet échapper à l'algorithme et le décideur doit utiliser son jugement pour les intégrer dans sa réflexion. Sans cela, il risque de déléguer implicitement la prise de décision aux concepteurs du système algorithmique (Oswald, 2018).

Le **biais d'adhésion sélective aux recommandations algorithmiques**, en revanche, contraste avec le biais d'automatisation. Ce biais d'adhésion sélective se manifeste par une propension des décideurs ou des utilisateurs à accepter ou à refuser les conseils algorithmiques, selon qu'ils confirment ou contredisent leurs préjugés préexistants. Bien qu'elle soit peu étudiée dans le contexte de l'aide à la décision algorithmique, cette problématique est souvent liée au traitement biaisé de l'information en administration publique. Certaines recherches montrent que ce biais n'est pas l'apanage des algorithmes d'aide à la décision et que l'IA ne les renforce pas nécessairement. Les analyses suggèrent plutôt que de tels mécanismes de traitement sélectif de l'information seraient probablement identiques avec des recommandations émises par des humains (Alon-Barkat et Busuioc, 2023). Par exemple, un juge ou un agent des forces de l'ordre pourraient plus facilement accepter une évaluation algorithmique prédisant un « risque élevé » pour une personne de couleur et un « faible risque » pour une personne blanche, car ce résultat conforte bien souvent leurs préjugés. Cette hypothèse rejoint de nombreuses études démontrant que les décisions des agents publics peuvent être influencées par des stéréotypes, menant à une discrimination envers les minorités et les groupes défavorisés (Jilke et Tummers, 2018 ; Pedersen et collb., 2018 ; Assouline et collab., 2022). Reste à savoir si

cette tendance est exacerbée par les outils algorithmiques ou si elle est similaire à l'adhésion sélective aux recommandations humaines. Cette distinction est importante à établir pour comprendre les mécanismes sous-jacents à ce biais et pour établir des stratégies visant à le prévenir, l'atténuer ou le supprimer (Alon-Barkat et Busuioc, 2023).

Le biais d'automatisation et l'adhésion sélective aux recommandations algorithmiques méritent une attention particulière dans le secteur public. Comme nous l'avons vu précédemment, le pouvoir discrétionnaire des agents de première ligne est indispensable au bon fonctionnement des organisations publiques. C'est grâce à leur jugement professionnel et à une certaine latitude décisionnelle que l'on évite de tomber dans un autoritarisme bureaucratique. Cependant, l'essor de l'IA menace ce pouvoir discrétionnaire en réduisant l'espace réservé au jugement professionnel. Les systèmes d'apprentissage automatique appliquent souvent un modèle statistique standardisé à toutes les décisions afin de garantir l'uniformité des résultats. Cette manière de faire réduit la prise en compte des particularités et du contexte propre à chaque situation dans la prise de décision. Cet enjeu est d'autant plus important que de nombreux domaines de l'administration publique nécessitent l'usage du jugement discrétionnaire. Certains estiment que l'introduction de systèmes d'aide à la décision automatisée pourrait soulever des questions juridiques, notamment si les algorithmes entravent le pouvoir discrétionnaire des agents publics (Cobbe, 2019). D'autres chercheurs estiment qu'il faudrait inverser les termes du problème et qu'il conviendrait de « créer un nouveau cadre où les algorithmes ne seraient pas uniquement utilisés pour générer les meilleures prédictions, mais conçus pour aider les humains à prendre de meilleures décisions dans des situations réelles » (Green et Chen, 2019, p. 90, trad.).

* * *

Après avoir parcouru les différentes étapes du pipeline de l'IA, il ressort clairement que des biais peuvent survenir à chacune de ces étapes et qu'ils peuvent prendre des formes variées. Les biais qui se manifestent à une étape donnée du cycle de l'IA peuvent affecter les étapes suivantes ou rétroagir sur les étapes précédentes du cycle. Ces rétroactions compliquent la détection et l'évaluation des biais potentiels, sans même parler de leur correction ou de leur élimination. Les conséquences de ces biais varient également selon les personnes ou les groupes qui sont concernés par les décisions administratives prises avec ou par ces systèmes algorithmiques. Nous avons également constaté que ces biais peuvent être introduits par divers acteurs, qu'il s'agisse des initiateurs, des concepteurs, des développeurs ou des utilisateurs des systèmes d'IA. Certains experts estiment que la combinaison de plusieurs facteurs que nous avons présentés jusqu'à présent risque d'aboutir à une administration déshumanisée, injuste, voire diabolique (*administrative evil*), si l'on se réfère au registre théorique mobilisé par les spécialistes de l'éthique publique américaine. Selon eux, l'utilisation de l'IA peut accroître les dérives bureaucratiques par l'interconnexion de plusieurs facteurs. Tout d'abord, l'IA introduit un biais d'automatisation lorsque les décisions algorithmiques sont souvent acceptées sans aucune remise en question préalable, favorisant ainsi une culture de la rationalité technologique qui peut négliger les considérations éthiques. De plus, le déploiement sous pression des systèmes d'IA, parfois sans tests adéquats, peut conduire à des solutions inappropriées pour des problèmes complexes. Ce phénomène s'accompagne d'un détournement des objectifs organisationnels où l'efficacité technologique prend le pas sur les missions de

service public. Par ailleurs, la complexité des systèmes d'IA est bien souvent synonyme d'opacité et d'inexplicabilité, ce qui rend difficile la compréhension des décisions et diminue la transparence. Ensemble, ces facteurs créent un environnement propice à des décisions administratives qui peuvent être nuisibles, même si elles sont prises dans un contexte qui à l'origine visait à améliorer l'efficacité et la qualité du secteur public (Young et collab., 2021).

L'énumération des biais décrits dans ce chapitre illustre le fait que ces problèmes ne sont pas seulement techniques, mais qu'ils reflètent aussi des choix et des préjugés qui influencent le développement et l'intégration de l'IA dans le secteur public. Avec les progrès rapides de l'IA et de ses technologies disruptives, il est important de rappeler que les types de biais ne sont pas figés dans le temps et qu'ils peuvent évoluer en fonction de l'utilisation des systèmes algorithmiques par les organisations publiques. De plus, de nouveaux risques, encore imprévus, pourraient émerger à mesure que la technologie évoluera ou que son utilisation au sein des organisations publiques se répandra.

CHAPITRE 3

Pistes d'action pour une administration numérique sans discrimination

Face aux risques de biais que nous avons présentés, certains pourraient suggérer de renoncer complètement à l'utilisation de l'IA dans le secteur public. Cependant, compte tenu de la rapidité des progrès technologiques en la matière et de l'utilisation croissante de l'IA dans le secteur public, il nous semble difficile de revenir en arrière et de renoncer à ces outils. Plutôt que de tenter de remettre « le génie dans la bouteille », nous pensons qu'il est nécessaire de trouver des moyens pour prévenir, réduire ou supprimer ces biais, afin de nous diriger vers une administration numérique sans discrimination.

1. LE RÔLE ESSENTIEL DES POUVOIRS PUBLICS DANS LA LUTTE CONTRE LES BIAIS ALGORITHMIQUES

Comme nous l'avons mentionné en introduction, le secteur public joue un rôle essentiel dans le développement et l'adoption de l'IA dans nos sociétés. Toutefois, les débats autour de la participation de l'État dans le domaine de l'IA se concentrent principalement sur les actions qu'il pourrait ou devrait envisager pour encadrer, soutenir ou promouvoir l'IA. Dans cette perspective, l'État est perçu comme le principal responsable de la création de conditions favorables

et de la mise en place des cadres balisant l'utilisation de l'IA par la population ou par les entreprises privées. Cette approche tend à rendre invisible, voire à négliger, le rôle de l'État en tant que principal acheteur et bénéficiaire direct de l'adoption et du déploiement de l'IA.

Cette situation met en lumière une tension importante relative aux biais algorithmiques : d'une part, les autorités publiques jouent un rôle de régulation face à une technologie potentiellement problématique, mais, d'autre part, en tant qu'utilisatrices, elles sont idéalement placées pour orienter et encourager le déploiement de ces outils dans de nombreux secteurs de la société. Bien que le développement des systèmes algorithmiques soit majoritairement le fait d'entreprises privées, l'État dispose d'une influence considérable sur l'acceptabilité sociale et l'utilisation responsable de ces technologies, grâce à son rôle de consommateur et d'utilisateur de l'IA. À cet égard, la *Stratégie québécoise d'introduction de l'IA dans l'administration publique (2021-2026)* contient plusieurs dispositions qui visent à positionner l'administration publique « comme [un] acteur exemplaire de l'IA » (Gouvernement du Québec, 2021, p. 2) et de garantir une utilisation conforme aux valeurs de l'administration publique « afin de protéger la population et [d']assurer sa confiance » (Gouvernement du Québec, 2021, p. 24).

1.1 L'IA et l'administration publique : une relation ambiguë

L'introduction et l'utilisation de l'IA dans le secteur public suscitent à la fois des espoirs et des craintes en ce qui concerne la discrimination. D'un côté, certains estiment que les algorithmes peuvent favoriser l'équité et la justice en appliquant des règles de manière impartiale, limitant ainsi les biais humains (Gaon et Stedman, 2019). Ces chercheurs

optimistes quant à l'équité et l'apprentissage automatique mentionnent que, pour que l'algorithme fonctionne comme prévu, il est nécessaire de formuler clairement et de manière explicite les intentions derrière les décisions. Par conséquent, ces chercheurs estiment que

> [l]a prise de décision fondée sur des données algorithmiques peut potentiellement être plus transparente que la prise de décision humaine. Elle nous oblige à formuler nos objectifs décisionnels et nous permet de comprendre clairement les compromis entre nos attentes. […] Une autre raison d'être optimiste, c'est que le passage à la prise de décision automatisée et à l'apprentissage automatique nous offre l'occasion de renouer avec les fondements moraux de l'équité. Les algorithmes nous obligent à être explicites quant au but du processus décisionnel. Et il est beaucoup plus difficile de dissimuler nos intentions imprécises ou cachées lorsque ces objectifs doivent être exprimés de manière formelle. Ainsi, l'apprentissage automatique nous incite à débattre plus efficacement de l'équité de différentes politiques et procédures de prise de décision (Barocas et collab., 2022, p. 21, trad.).

D'autres chercheurs rappellent que les biais détectés dans les algorithmes peuvent théoriquement être corrigés en ajustant la programmation, notamment lors des phases de modélisation et de validation d'un système. Cependant, des critiques soulignent quant à eux que les algorithmes peuvent refléter les biais des concepteurs et des données utilisées pour l'entraînement des algorithmes d'apprentissage automatique. Ces biais pourraient rester indécelables avant le déploiement des systèmes. L'utilisation de l'IA aurait pour conséquence de désavantager certains groupes sociaux et d'engendrer des conséquences négatives sur la fourniture des services publics à ces groupes. Outre les risques de discrimination, des préoccupations émergent quant à l'utilisation de l'IA à des fins de profilage racial ou de surveillance, soulevant des questions relatives à l'éthique et à la protection de la vie privée (Desouza et collab., 2020 ; Newman et Mintrom, 2023).

Les défenseurs et les détracteurs de l'IA s'affrontent en mettant en lumière les aspects positifs ou négatifs de cette technologie. Chaque camp présente des arguments pertinents qui nourrissent le débat actuel sur l'intégration de l'IA dans les organisations publiques. Toutefois, ce clivage occulte une dimension fondamentale : l'incertitude inhérente à cette technologie. En reconnaissant cette incertitude, il devient possible de mieux comprendre les risques liés à l'IA et de les voir comme le fruit d'une série de décisions et de compromis effectués tout au long du pipeline de l'IA. Dès lors, certains estiment qu'une élimination complète des biais pourrait être contre-productive, car elle risquerait de présenter aux utilisateurs une vision déformée de la réalité, voire de rendre plus difficile la détection des biais réellement présents dans la société. Ils ajoutent que, s'ils sont bien gérés, les biais non seulement seraient inoffensifs, mais pourraient être utiles à l'utilisateur en lui fournissant une information sur ces biais. Dès lors, selon les tenants de cette perspective, l'objectif ne devrait pas être d'éliminer systématiquement les biais de manière *a priori*, mais plutôt de les identifier, de les mesurer et de les traiter comme une caractéristique du système d'IA laissant aux utilisateurs le choix de les ajuster ou non en fonction de leurs besoins (Demartini et collab., 2021).

En effet, il est essentiel de comprendre que certains biais algorithmiques ne pourront probablement jamais être totalement éliminés et qu'il faudra une gestion et une vigilance continues pour en limiter les effets négatifs. Cette réalité est en partie due à la nature sociale des biais que nous avons représentée par l'écart 1 dans la figure 1. Les biais sociaux ne sont pas statistiques ou établis une fois pour toutes. Certains biais s'estompent, tandis que d'autres émergent au rythme de l'évolution des valeurs et de la société. Même si l'on imagine un système d'IA parfaitement en phase avec les

valeurs actuelles de la société, certaines de ces valeurs pourraient encore être biaisées ou des discriminations aujourd'hui invisibles pourraient être révélées dans le futur. Cette mise en lumière nécessiterait une adaptation des valeurs sociales et un ajustement des algorithmes en conséquence.

De plus, la recherche de solutions ne doit pas négliger l'évolution rapide de la technologie. En effet, cette évolution rapide permettrait de réduire, voire de résoudre complètement les problèmes associés aux biais algorithmiques, mais elle pourrait aussi en changer la nature ou les accélérer. Lorsqu'on travaille à résoudre un problème, il est tentant de le cristalliser dans son état actuel et d'essayer de le résoudre, sans prendre en considération toutes ses facettes. Compte tenu de ces deux réalités – l'évolution constante de la société et de ses valeurs, d'une part, et l'évolution rapide des technologies, d'autre part –, il existe un risque que le processus de résolution d'un problème aboutisse à une solution inadaptée. En effet, le problème peut avoir disparu, s'être transformé ou complexifié, au point que la solution proposée ne soit plus pertinente.

1.2 Prendre conscience des risques de l'IA

Des chercheurs ont mis en lumière une série de risques majeurs que l'IA pose à l'humain, aux organisations et à la société. En étant conscient de l'existence de ces risques, il est ensuite possible de les réduire et de chercher un équilibre permettant de tirer le meilleur parti de l'IA (Bannister et Connolly, 2020). Nous présentons ces risques ici pour qu'ils alimentent la réflexion des responsables politiques et des gestionnaires publics lorsqu'ils envisagent d'intégrer l'IA dans le secteur public.

Le premier risque concerne une **dépendance excessive à la technologie**, qui pourrait nuire à l'autonomie des individus et affaiblir les compétences humaines. Si l'automatisation permet d'éliminer les tâches répétitives et d'optimiser la gestion de notre temps, elle soulève des enjeux qui vont au-delà de la simple imitation des capacités humaines. En effet, la rapidité des transformations pourrait, à plus long terme, nuire au développement des êtres humains et à leurs capacités individuelles à s'adapter à leur nouvel environnement de travail. Pour la société, l'affaiblissement des compétences humaines pourrait engendrer des vulnérabilités importantes, surtout si les systèmes d'IA tombent en panne et que les êtres humains ont perdu certaines compétences en comptant trop sur ces systèmes pour accomplir leur travail.

Le deuxième risque est celui de la **déresponsabilisation et du désengagement moral**. Ce risque pourrait se manifester si les individus ou les organisations s'appuient trop sur l'IA pour prendre des décisions, diminuant ainsi leur propre implication et leur sens de la responsabilité. Les décisions prises par des algorithmes impliquent l'intégration de valeurs, que ce soit directement dans le code ou indirectement par les données utilisées ou les choix effectués lors du développement du système d'IA. Lorsque des systèmes algorithmiques prennent des décisions, il est donc essentiel que l'humain reste responsable des choix qui sont faits, même si ceux-ci sont assistés par l'IA. Si les décisions sont déléguées à des systèmes d'IA, sans une réflexion préalable suffisante sur les valeurs intégrées dans ces systèmes, il devient facile de rejeter la responsabilité des erreurs sur la technologie. Cette attitude peut affaiblir la responsabilité morale sur les plans tant individuel qu'organisationnel. À l'inverse, certains estiment qu'une IA bien conçue pourrait non seulement améliorer l'autonomie individuelle, mais aussi renforcer nos systèmes

moraux. Contrairement à l'idée selon laquelle une plus grande autonomie des systèmes d'IA diminuerait proportionnellement celle des humains, la réalité est plus complexe. En fait, l'IA utilisée correctement pourrait accroître la qualité des décisions humaines en enrichissant la réflexion, plutôt qu'en se substituant à l'humain. Par exemple, un système d'aide au diagnostic médical peut enrichir le processus décisionnel d'un médecin, sans le remplacer. Sur le plan éthique, déléguer certaines décisions à l'IA ne réduit pas nécessairement la responsabilité morale des humains, mais peut au contraire la renforcer, en obligeant les utilisateurs à réfléchir aux valeurs qui doivent être intégrées dans ces systèmes.

Le troisième risque appréhende la **perte de contrôle collectif** face aux capacités croissantes des systèmes d'IA dans la société. Les systèmes d'IA offrent de nombreuses occasions d'améliorer les capacités collectives humaines en contribuant à trouver des solutions à des problèmes majeurs qui affectent la société. Cependant, s'appuyer excessivement sur ces systèmes pour renforcer nos capacités pourrait nous amener à déléguer des décisions importantes à des machines, alors que celles-ci devraient être prises par des humains. Ces risques peuvent augmenter à mesure que l'IA prendra en charge davantage de fonctions et que des composantes, voire l'intégralité, du pouvoir décisionnel individuel ou organisationnel lui seront déléguées. Il est donc essentiel de trouver un équilibre entre l'exploitation du potentiel de l'IA et le maintien d'un contrôle humain sur son développement et son utilisation.

Face à ces risques, et en raison de l'intégration croissante de systèmes d'IA dans le secteur public, plusieurs pays ont réaffirmé que la responsabilité des décisions incombe en dernier ressort aux organisations et aux agents publics. Il s'agit d'un élément important dans la gouvernance des systèmes

d'IA, en particulier dans des situations où la prise de décision automatisée pourrait affecter les droits des personnes ou l'accès aux services publics. Les droits garantis par ces lois et règlements, tels que le droit à une intervention humaine et le droit de contester les décisions automatisées, sont essentiels pour préserver l'équité des processus décisionnels automatisés (Besse et collab., 2017). Par ailleurs, de nombreuses organisations de surveillance ou de contrôle insistent pour qu'une supervision humaine soit maintenue dans le processus décisionnel automatisé. C'est par exemple le cas en France où les magistrats du Conseil d'État, dans une étude commandée par le premier ministre, ont rappelé que « l'IA publique de confiance doit […] être fondée sur le principe de primauté humaine » (Conseil d'État, 2022, p. 47). Cette supervision humaine ne se limite toutefois pas à une simple validation des résultats de l'algorithme, mais implique une évaluation approfondie de son influence sur les décisions qui sont prises. Cela signifie qu'il est nécessaire d'établir des protocoles d'évaluation *a posteriori* pour mesurer l'effet réel des algorithmes sur les décisions finales et s'assurer qu'elles ne sont pas contraires aux principes constitutionnels, au cadre légal ou réglementaire en vigueur ou à la volonté générale de la population (Cluzel-Métayer, 2018 ; De-Arteaga, 2020).

1.3 Un encadrement de l'IA dans le secteur public à renforcer

À l'heure actuelle, il existe peu de cadres généraux permettant d'encadrer l'utilisation de l'IA au sein des organisations publiques. Les exemples qui existent se concentrent principalement sur une approche d'évaluation d'impact des risques qui a pour ambition de « dérisquer » les modèles d'IA. C'est par exemple le cas de la *Directive sur la prise de décisions*

automatisée du gouvernement fédéral canadien, adoptée en 2019. Selon cette approche, les mesures à prendre lors du développement et de l'utilisation d'un système d'IA dans le secteur public sont directement liées à la nature des risques qu'il présente : plus le risque est élevé, plus les exigences et les mesures à mettre en place sont importantes. Cela reflète le principe de proportionnalité que l'on retrouve dans de nombreux énoncés promouvant l'éthique de l'IA, telle la *Recommandation sur l'éthique de l'intelligence artificielle* de l'UNESCO adoptée, en 2021, par 193 États. Cette approche sur la gestion des risques relatifs à l'IA prévaut dans de nombreux pays. Nous la retrouvons dans le *Règlement européen sur l'intelligence artificielle* de 2024 (*AI Act*), qui formule des exigences pour garantir que les systèmes d'IA sont sûrs, fiables et transparents.

En parallèle, en l'absence d'un cadre global balisant l'utilisation de l'IA dans le secteur public, plusieurs lois spécifiques continuent de réglementer ce que les organisations publiques peuvent faire ou non en la matière. Par exemple, des lois comme le *Règlement général sur la protection des données* (RGPD) en Europe ou la *Loi modernisant des dispositions législatives en matière de protection des renseignements personnels* au Québec (communément appelée loi 25) influencent l'utilisation de l'IA en régulant la collecte, l'utilisation et la protection des données et des renseignements personnels. Ces lois balisent partiellement l'utilisation de l'IA, mais ne remplacent pas un cadre général sur une utilisation responsable de l'IA au sein du secteur public.

L'utilisation de l'IA dans le secteur public doit être abordée avec prudence. La technologie ne doit pas éclipser le fait que les organisations publiques sont soumises au droit. L'IA est un outil au service de l'accomplissement des missions

d'intérêt général de l'État, et non une finalité en soi. L'IA facilite des tâches, telles que l'attribution de prestations sociales, la lutte contre la fraude fiscale, la fourniture de soins de santé ou encore la protection de groupes vulnérables. Les algorithmes d'aide à la décision sont eux aussi soumis aux exigences juridiques qui encadrent toute autre décision publique, en matière notamment de transparence et de responsabilité. Les agents publics et les responsables politiques doivent rendre compte de leurs décisions, qu'elles soient prises par une personne ou assistées, voire formulées, par un système d'IA (Waller et Waller, 2020). De leur côté, les citoyennes et les citoyens

> ont le droit de contester et de demander réparation pour toute décision qui les affecte négativement. Lorsque les répercussions des décisions prises par les algorithmes ont une incidence sur la vie et le bien-être des personnes, il est essentiel de pouvoir faire pleinement confiance à l'algorithme. Les conséquences de mauvaises décisions peuvent être graves, comme cela a été démontré dans le domaine de la justice pénale et des services de protection de l'enfance (Waller et Waller, 2020, p. 4, trad.).

Les effets néfastes des biais algorithmiques se manifestent d'abord sur le plan individuel, mais ils peuvent aussi s'étendre à l'ensemble des membres d'un groupe ou d'une communauté marginalisée, renforçant ainsi les inégalités et les stéréotypes que subissent ces groupes.

Jusqu'à présent, la recherche en apprentissage automatique axée sur l'équité (*fair ML*) s'est concentrée principalement sur la conception d'algorithmes et de modèles d'apprentissage plus justes, en traitant l'équité comme une propriété interne du système algorithmique (Selbst et collab., 2019). Ce défi est complexe, car il est difficile de parvenir à une définition claire et universellement acceptée de l'équité, comme nous l'avons indiqué précédemment. Les développeurs d'IA doivent donc naviguer entre de nombreuses

interprétations possibles du concept d'équité qu'ils tentent ensuite de traduire en langage informatique. Pour cela, ils essaient de définir les meilleures approximations possibles de l'équité en tenant compte des contraintes techniques ou des indicateurs spécifiques qui traduisent et opérationnalisent le concept. De plus, la majorité de ces travaux en apprentissage automatique axés sur l'équité se limitent bien souvent à chercher à améliorer les paramètres du système d'IA, en se concentrant sur le modèle d'apprentissage automatique, ses données d'entrée et de sortie, sans nécessairement s'intéresser aux influences externes ou au contexte plus large dans lequel ces systèmes d'IA sont utilisés. Des experts expliquent que les conséquences de cette abstraction du contexte social

> rendent les interventions techniques proposées inefficaces, inexactes et parfois dangereusement erronées lorsqu'on les applique dans la société. [De plus,] en faisant abstraction du contexte dans lequel ces systèmes seront déployés, les chercheurs sur l'équité en apprentissage automatique ne tiennent pas compte du contexte plus large, y compris des informations nécessaires pour créer des résultats plus équitables, ou même pour comprendre le concept d'équité en tant que tel (Selbst et collab., 2019, p. 59, trad.).

Ces experts illustrent cette inadéquation par cinq pièges dans lesquels les promoteurs des travaux sur l'équité en apprentissage automatique peuvent tomber, même lorsqu'ils cherchent à être plus attentifs au contexte sociétal que la science des données traditionnelle. Ces pièges sont décrits dans l'encadré 5.

Chacun de ces pièges découle d'une absence de prise en compte de l'interdépendance entre le contexte social et la technologie, ainsi que du besoin d'une compréhension approfondie des dynamiques sociales pour formuler des solutions adaptées aux problèmes rencontrés. En effet, il est important de mettre en garde contre le danger de se concentrer

uniquement sur la correction des biais sous l'angle technique ou méthodologique, car nous risquerions alors de ne pas porter suffisamment d'attention aux aspects plus larges, tels que la pertinence et la nécessité de l'utilisation d'un système algorithmique prédictif dans un contexte administratif particulier (Babuta et Oswald, 2019).

ENCADRÉ 5
Pièges à éviter

- **Piège du solutionnisme technologique** : en estimant que la technologie offre une réponse à tout, ne pas envisager que la meilleure solution à un problème pourrait être de se passer de la technologie.
- **Piège du cadrage** : ignorer ou ne pas reconnaître les limitations imposées par le choix d'un cadre conceptuel particulier sur la nature des résultats générés par l'algorithme, ce qui peut influencer l'interprétation des résultats et la prise de décision basée sur ces résultats.
- **Piège du formalisme** : incapacité à rendre pleinement compte de la complexité et de la variété des définitions de concepts sociaux qui ne se prêtent pas toujours à des formalisations mathématiques.
- **Piège de l'effet d'entraînement** : ne pas comprendre comment l'introduction d'une technologie dans un système social existant peut transformer les comportements et les valeurs propres à ce système.
- **Piège de la transférabilité** : incapacité à comprendre comment la transposition des solutions algorithmiques conçues pour un contexte social spécifique peut être trompeuse, inexacte ou nuisible lorsque ces solutions sont exportées et appliquées dans un autre contexte.

Source : Selbst et collab., 2019, p. 60-63, trad.

2. LES QUESTIONS À SE POSER POUR ÉVITER LES BIAIS ALGORITHMIQUES DANS LE SECTEUR PUBLIC

2.1 L'utilisation de l'IA est-elle nécessaire ?

Dès le moment où une organisation publique envisage de recourir à l'IA ou à des systèmes de décision automatisée, il est essentiel d'évaluer la pertinence de cette utilisation. L'IA doit être considérée comme un moyen, et non comme une fin en soi. À cet égard, même en concevant l'IA ou le système algorithmique comme un simple outil, il est important d'explorer les options qui permettent de résoudre un problème ou d'accomplir une tâche sans se limiter aux solutions technologiques, au risque de tomber dans le piège du solutionnisme technologique (Selbst et collab., 2019).

Dans le secteur public, la réflexion sur l'utilisation de l'IA ou des systèmes algorithmiques doit être guidée avant tout par les principes relatifs à la poursuite de l'intérêt général et la construction du bien commun. Il s'agit donc d'évaluer si ces technologies servent véritablement les besoins collectifs et apportent une plus-value aux organisations du secteur public. Cette réflexion nécessite de porter une attention particulière au contexte dans lequel les algorithmes seront développés et déployés, ainsi qu'à l'évaluation du rapport coûts-bénéfices pour la collectivité, en ce qui concerne la fourniture de services et la relation entre l'État et la population. En cela, cette réflexion est intimement liée à la question de l'acceptabilité sociale de l'IA par les agents publics et par la population.

D'autres critères peuvent ensuite être utilisés pour évaluer la pertinence d'un algorithme. Il s'agit de la bienfaisance (les bénéfices générés), la non-malfaisance (l'absence de préjudice), l'autonomie (le respect du libre arbitre humain), la justice (légalité, l'équité et l'honnêteté) et l'explicabilité (la capacité à comprendre le fonctionnement et le

raisonnement du processus décisionnel de l'algorithme). Il est également nécessaire d'évaluer si les résultats d'un algorithme seront à la fois fiables et compréhensibles pour toutes les parties prenantes (Waller et Waller, 2020). En effet, certaines personnes pourraient ne pas comprendre ou ne pas obtenir les informations nécessaires sur le fonctionnement du système d'IA en raison de spécifications techniques inappropriées dont nous avons parlé dans la section sur l'opacité des systèmes d'IA.

En raison de la complexité des systèmes d'IA, la plupart des membres de la population ne disposent pas des connaissances techniques suffisantes et n'ont pas la possibilité de consulter des experts pour obtenir les informations ou les explications nécessaires. Dans ce cas, bien que les principes de transparence, de responsabilité ou d'explicabilité soient théoriquement envisageables, leur mise en pratique peut s'avérer difficile. Les recherches sur l'IA explicable visent à développer des langages, des processus et des rapports capables de rendre accessibles et compréhensibles au plus grand nombre de personnes le fonctionnement et les résultats générés par des algorithmes d'apprentissage automatique. Dans la pratique, il sera cependant difficile d'atteindre une explicabilité complète pour le grand public, ce qui complique l'évaluation de l'acceptabilité sociale et de la pertinence de l'utilisation de l'IA et des systèmes algorithmiques dans le secteur public. Cette difficulté ne doit pas servir d'excuse pour relâcher les efforts en la matière, car il est essentiel de garantir le principe de transparence afin d'assurer le respect des lois, notamment en cas de problèmes, de plaintes ou d'enquêtes.

Si l'utilisation de l'IA est considérée comme pertinente au sein d'une organisation publique, il est essentiel d'assurer la transparence quant à cet usage. À cet égard, plusieurs

pays imposent aux administrations publiques d'informer les usagers lorsque l'IA est utilisée dans le traitement de leurs dossiers. En Europe, le *Règlement général sur la protection des données* (RGPD) consacre cette obligation. Au Québec, l'*Énoncé de principes pour une utilisation responsable de l'intelligence artificielle par les organismes publics*, adopté en 2024, indique que les organisations publiques doivent informer les citoyennes et les citoyens ainsi que les entreprises de la nature, de la portée et du moment où les systèmes d'IA sont utilisés. Le ministère de la Cybersécurité et du Numérique propose à cet effet aux ministères et organismes publics québécois d'adopter un « visuel qui permet de confirmer aux usagers finaux que le service qu'ils reçoivent est généré par un système d'IA » (Arrêté ministériel nº 2024-02, p. 5453).

En résumé, pour éviter les biais algorithmiques, il faut tout d'abord s'assurer de l'acceptabilité sociale de l'utilisation de l'IA dans le secteur public et évaluer la pertinence de l'intégration d'un outil d'aide à la décision automatisée dans le fonctionnement des organisations publiques. Cette évaluation nécessite un niveau suffisant de transparence et d'explicabilité pour que les parties prenantes puissent comprendre le fonctionnement du système, tout en adoptant une approche globale qui intègre non seulement les aspects techniques, mais aussi les implications éthiques, sociales et contextuelles de l'utilisation de systèmes d'IA dans les organisations publiques.

2.2 Les données utilisées par le système d'IA sont-elles de qualité ?

Nous avons souligné à plusieurs reprises dans cet ouvrage que la qualité des données est primordiale dans la conception et l'utilisation de l'IA. Si les systèmes d'aide à la décision

sont formés à partir de données de mauvaise qualité, ils produiront inévitablement des résultats de qualité médiocre qui risquent de générer, de reproduire ou d'amplifier les biais dans la société.

Évaluer la qualité des données afin de prévenir, de réduire ou d'éliminer les biais algorithmiques dans le secteur public représente un défi qui pourra être surmonté en prenant en considération quatre éléments, entre autres.

Le premier élément est la **disponibilité des données**. Les données sont pour l'IA ce que l'essence est pour un moteur à explosion : un élément indispensable à son fonctionnement. Les entreprises technologiques l'ont bien compris et ont investi, depuis plusieurs années, des moyens considérables pour collecter et organiser les données afin qu'elles puissent alimenter leurs solutions d'IA. Par exemple, à l'aide de son moteur de recherche, du site YouTube et de l'application de navigation GPS (Maps), Google collecte des données massives sur les utilisateurs en ligne sur Internet et hors ligne (comme les visites en magasin) par la géolocalisation et les applications qui partagent ces informations avec l'entreprise de la Silicon Valley. Ces données sont utilisées pour améliorer les algorithmes d'IA de Google, en particulier lors de la diffusion de publicité ciblée.

Les administrations publiques ne disposent quant à elles pas toujours de données de qualité suffisante pour développer des systèmes d'IA. Ce constat peut sembler surprenant si l'on songe à la quantité de données dont disposent certains ministères. Malgré ce volume de données, celles-ci ne sont pas toujours utiles ou exploitables pour l'IA, car elles peuvent être réparties entre différents services au sein d'un même ministère, être mal structurées, incomplètes ou non standardisées, ce qui rend difficile leur intégration dans des systèmes d'IA.

Le développement d'une infrastructure publique pour la gestion des grands ensembles de données est donc une étape indispensable au développement d'une IA de qualité. Ce défi est de taille et, pour le surmonter, il sera sans doute nécessaire d'envisager un partenariat avec des entreprises du secteur privé, à moins de développer rapidement une expertise interne chargée de cette mission. Les chercheurs qui plaident en faveur de la collaboration avec les entreprises du secteur privé estiment que ce partenariat permettra d'accélérer la création d'ensembles de données représentatives qui sont essentiels pour assurer l'équité dans les modèles d'IA. Dans le domaine de la santé, par exemple, ces ensembles de données pourraient être mis à la disposition des entreprises qui développent des algorithmes en assurant la protection de la vie privée des patients et la répartition équitable des bénéfices générés au profit à la fois des organisations du secteur public, qui soignent les patients et produisent des données, et des entreprises du secteur privé qui ont l'expertise nécessaire pour utiliser les données et concevoir des algorithmes (Panch et collab., 2019).

Une autre piste à envisager pour améliorer l'accès aux données est l'ouverture des ensembles de données gouvernementales. L'établissement d'un cadre réglementaire balisant la gouvernance de ces données est nécessaire pour encadrer ce partage des données. Ce cadre devrait inclure des directives garantissant la compatibilité entre différents systèmes et l'interopérabilité des données, facilitant ainsi leur utilisation par divers acteurs ou organisations. L'objectif est de rendre ces données plus accessibles, tout en préservant leur intégrité et leur utilité pour une large gamme d'applications (Lloyd et Hamilton, 2018).

Le deuxième élément est la **préparation adéquate des données** d'entrée. Dans la majorité des ensembles de

données, il est fréquent de trouver des valeurs manquantes, ce qui nécessite l'utilisation de méthodes statistiques lors du prétraitement des données. Lorsque plusieurs ensembles de données sont utilisés ou combinés, il est essentiel de vérifier que les définitions, les indicateurs et les mesures des données fusionnées sont similaires afin d'éviter de comparer des pommes avec des poires. Pour la modélisation, certaines données, telles que les textes, doivent être converties en format numérique, car les modèles ne peuvent pas toujours traiter directement des données textuelles. Par exemple, les données géographiques, comme les codes postaux, doivent être reformatées pour être exploitables par les algorithmes. Les données peuvent être représentées numériquement en attribuant des valeurs chiffrées à des éléments, tels que le nombre d'universités dans un pays, ou des indicateurs, comme le revenu moyen ou le niveau de précarité. D'autres données plus qualitatives, comme les catégories, peuvent être converties sous forme d'étiquettes où chaque catégorie se voit assigner un code ou une étiquette spécifique facilitant leur traitement par les systèmes d'IA (Waller et Waller, 2020).

Le troisième élément est de veiller à la **représentativité des données**. Comme nous l'avons vu plus tôt, il est important que les données utilisées par les modèles d'IA reflètent les caractéristiques de la population dans son ensemble. Dans le cas contraire, il y a une forte probabilité que les décisions automatisées avantagent ou désavantagent certaines personnes ou certains groupes, en particulier les plus vulnérables (Cluzel-Métayer, 2020 ; Udoh, 2020). Des chercheurs recommandent de prioriser la collecte de données diversifiées, notamment celles qui concernent le genre, l'origine ethnique et la nationalité. Ils proposent également de recueillir des informations sur des facteurs tels que le statut d'incapacité, le statut socioéconomique, le

milieu de résidence (urbain ou rural), la langue maternelle ou le statut d'apprenant de langue seconde et le niveau de scolarité des parents (Baker et Hawn, 2022). Cette diversité dans la collecte de données permettrait de construire des modèles plus représentatifs de la société dans son ensemble. Bien que cette accumulation de données contribue à une meilleure représentativité, il ne faut pas oublier qu'elle implique la collecte d'une quantité importante de données personnelles sensibles. Il est donc essentiel d'ajuster cette collecte en fonction des besoins spécifiques des organisations publiques afin de réduire les risques liés à la confidentialité et à la protection des renseignements personnels.

Une étude réalisée dans le domaine de la santé illustre le fait que l'utilisation de données qui contiennent des informations relatives aux caractéristiques des patients permet de réduire le risque de discrimination algorithmique. À cet égard, plus le nombre de caractéristiques et d'informations disponibles est élevé, plus les développeurs sont en mesure d'entraîner un algorithme à reconnaître des tendances cachées, à apprendre de valeurs aberrantes et à s'adapter à un plus grand nombre de sous-ensembles d'individus ou de groupes de patients (Henderson et collab., 2022). C'est pour cette raison que les applications d'IA relatives à la santé se développent rapidement, en utilisant principalement des caractéristiques communes aux patients, telles que les symptômes, les résultats de tests, l'âge, le sexe ou l'origine ethnique. Un accès à des données robustes et diversifiées permet également d'éviter de sur-ajuster les petits ensembles de données qui peuvent manquer de représentativité ou de sous-ajuster les ensembles de données incomplets. Finalement les auteurs de cette étude insistent sur le fait que les données doivent être mieux recueillies auprès des populations vulnérables et marginalisées et que la communauté de recherche

dans le domaine de la santé doit continuer à être sensibilisée aux enjeux relatifs à la diversité. Que ce soit dans le domaine de la santé ou dans d'autres secteurs d'intervention publique, la prévention, l'atténuation et la suppression des biais algorithmique nécessitent une approche multidimensionnelle centrée sur la diversité et la représentativité des données utilisées (Lloyd et Hamilton, 2018).

L'État a un rôle à jouer en cette matière et il doit s'assurer que les développeurs utilisent des données diversifiées et représentatives dans la conception de systèmes d'IA qui seront déployés dans le secteur public. Pour y parvenir, le gouvernement et les organisations publiques devraient engager un dialogue avec les développeurs d'IA et établir une politique claire concernant les normes de données utilisées par les organisations publiques. Ces normes viseraient à garantir une représentation diversifiée dans les ensembles de données, afin de prévenir l'amplification des biais, et à veiller à ce que les systèmes d'IA soient conçus en intégrant le principe d'inclusion. Ces normes non seulement permettraient de protéger la population des préjudices potentiels, mais éviteraient également aux organisations publiques d'être aux prises avec des problèmes ou des scandales algorithmiques qui ternissent leur réputation (Lloyd et Hamilton, 2018).

Si des biais sont repérés, l'application de techniques dites de « débiaisage » permet l'ajustement des ensembles de données et la modification des algorithmes afin de les réduire (Besse et collab., 2017). Bien que ces méthodes puissent aider à atténuer certains biais, ces solutions techniques ne permettent pas de rendre pleinement compte des dimensions sociales ou culturelles qui contribuent à alimenter ou générer ces biais. Par conséquent, se concentrer uniquement sur des solutions techniques pourrait conduire à une vision simpliste du problème.

Le quatrième élément consiste à **reconnaître que les données ont des limites**, quels que soient les efforts déployés pour veiller à leur qualité. Il est essentiel de reconnaître explicitement ces limites et d'évaluer les répercussions que ces limitations peuvent avoir sur le fonctionnement du système algorithmique, afin de s'assurer que celui-ci demeure malgré tout pertinent et capable d'atteindre ses objectifs (Suresh et Guttag, 2021). Par exemple, il est important de différencier les biais liés à la sélection des données des biais historiques. Tandis que le premier biais peut être atténué par l'amélioration des méthodes et du système de collecte de données, le second persiste même avec des mesures parfaites, car les données reflètent des réalités passées considérées comme indésirables selon les valeurs actuelles de la société. Ces données sont donc biaisées, dans la mesure où les prédictions que l'algorithme génère auront tendance à reproduire ces anciennes réalités. Le fait d'opter pour des données historiques peut d'ailleurs être perçu comme un choix de nature politique, puisqu'il traduit une intention de s'appuyer sur une vision du passé pour atteindre des objectifs actuels (Levy et collab., 2021).

D'autres enjeux politiques doivent également être pris en compte à ce stade. Certains chercheurs, qui insistent sur la vérification et la validation des ensembles de données avant leur utilisation, remettent en question l'idée de neutralité dans la gestion des biais. Ils estiment que la suppression des caractéristiques sensibles, telles que l'origine ethnique, le genre, l'âge ou la religion, et le nettoyage des données pour supprimer ou corriger les éléments biaisés de manière à rendre les données « neutres » peuvent s'avérer problématiques et insuffisants pour atténuer ou éliminer les biais algorithmiques (Lloyd et Hamilton, 2018). Les questions soulevées à ce sujet visent notamment les sujets suivants :

par qui et de quelle manière sont définies les informations ou caractéristiques sensibles à supprimer ? Si la prise en compte des situations et caractéristiques personnelles est essentielle pour un traitement équitable, l'utilisation de données « neutres » permet-elle réellement d'assurer un traitement algorithmique juste ? La correction ou la suppression de sources de biais potentielles ou avérées risque-t-elle de masquer ou d'invisibiliser certaines réalités ou certains phénomènes importants à considérer pour progresser vers une société plus équitable ?

Masquer ou protéger certaines variables ne suffira donc pas à réduire ou à supprimer tous les risques de biais. Une fois de plus, nous constatons que les biais algorithmiques ne sont pas liés uniquement à la technique employée et qu'une approche strictement technique ne peut à elle seule les résoudre. C'est pour cette raison que des chercheurs recommandent la mise en place, dès les premières étapes du développement et tout au long du pipeline de l'IA, d'équipes multidisciplinaires en science des données, sensibilisées aux particularités du contexte d'application des algorithmes, ainsi qu'un encadrement réglementaire et une évaluation adéquate des algorithmes (Panch et collab., 2019).

2.3 Comment l'algorithme fonctionne-t-il ?

Sans une compréhension adéquate du fonctionnement interne des algorithmes, il est difficile de s'assurer de leur fiabilité et de réduire ou corriger les biais. Sans transparence et explicabilité, la confiance du public envers ces technologies pourrait s'éroder à mesure que les risques de discrimination dans l'administration publique pourraient survenir ou s'aggraver.

Les discussions sur la transparence de l'utilisation de l'IA par les organisations publiques se sont intensifiées ces dernières années. Ces réflexions soulignent la nécessité de fournir des informations détaillées et accessibles concernant l'utilisation de l'IA dans les organisations publiques. De nombreux spécialistes plaident pour que l'État fournisse ces informations d'une manière proactive, c'est-à-dire sans attendre une demande explicite par des procédures d'accès à l'information ou une injonction imposée par un juge. Un mécanisme qui gagne en popularité à l'échelle internationale est la création de registres publics répertoriant les algorithmes utilisés dans le secteur public. Ces registres permettent de fournir une information détaillée sur chaque système d'IA comme nous pouvons le voir dans l'encadré 6. Un recensement récent de ces registres à l'échelle internationale a permis de repérer près de 70 registres accessibles aux différents ordres de gouvernement, que ce soit aux niveaux central, régional ou local. « Les États-Unis comptent 37 registres actifs, suivis des Pays-Bas avec 14, de la France avec 4 et de la Colombie avec 3 » (GPAI, 2024, p. 14, trad.).

Comprendre le fonctionnement des algorithmes est un défi essentiel à relever pour assurer une utilisation responsable de l'IA. Ce défi est d'autant plus grand qu'il est souvent difficile de saisir clairement le fonctionnement des algorithmes et ce qui se passe au cœur de la « boîte noire ». De plus, cette compréhension du fonctionnement des systèmes d'IA est souvent entravée par la volonté des entreprises de protéger leurs secrets commerciaux et leur propriété intellectuelle (Besse et collab., 2021). En cette matière, il est nécessaire de trouver un équilibre entre la transparence et la protection des droits de propriété intellectuelle. Une solution envisagée est de permettre à des scientifiques accrédités d'avoir accès aux algorithmes pour en effectuer une analyse approfondie

similaire au travail des agences chargées d'examiner l'efficacité des nouveaux médicaments ou des dispositifs médicaux avant leur disponibilité sur le marché (Gerke et collab., 2020). D'autres chercheurs encouragent les responsables politiques à établir un cadre de gouvernance de l'IA afin de créer des agences régulatrices. Ces agences seraient responsables d'établir des standards garantissant une utilisation responsable des algorithmes, d'approuver et de surveiller la conception et le déploiement des algorithmes, surtout de ceux qui peuvent avoir des répercussions importantes sur la vie des personnes ou l'évolution de la société (Bannister et Connolly, 2020).

ENCADRÉ 6
Informations disponibles dans un registre public des algorithmes

1. Nom du système.
2. Objectifs, tâches et résultats du système (description de ce que fait le système et/ou comment il le fait).
3. Unité chargée du déploiement du système (ou unité responsable).
4. Architecture du système, techniques et méthodes utilisées pour construire le système.
5. Informations sur l'état du système et son processus de développement (par exemple, prototype, phase d'essai, production, mise à l'arrêt ou retrait).
6. Sources de données et indication de l'utilisation de données personnelles, si nécessaire.
7. Informations sur le développement du système (par exemple, en interne ou avec des développeurs externes).
8. Informations de contact de l'unité et/ou de l'agent public responsable du déploiement du système.
9. Informations sur le code source.

Source : GPAI, 2024, p. 16-17, trad.

La question de l'explicabilité des algorithmes est très complexe, car elle comprend des aspects relatifs à la transparence, à la responsabilité et à la compréhension des processus décisionnels automatisés. Certains experts considèrent qu'une décision algorithmique est jugée explicable lorsqu'il est possible de rendre compte de son fonctionnement de manière explicite en se basant sur des données et des caractéristiques bien définies dans le contexte de la situation posée (Besse et collab. 2021). Autrement dit, l'explicabilité suppose la capacité de démontrer clairement la relation entre les valeurs de certaines variables (les caractéristiques) et leur influence sur les prédictions ou les décisions du modèle algorithmique. Dans le cas d'une organisation publique, cela nécessite la mise en place de mécanismes qui permettent aux agents publics de bien comprendre le fonctionnement des algorithmes d'aide à la décision, les données qu'ils utilisent et la logique sous-jacente à la prise de décision. Cette information devrait également être communiquée de manière compréhensible au public afin de garantir la transparence et de renforcer la confiance de la population envers les organisations publiques qui utilisent l'IA.

Plusieurs méthodes ont été proposées pour améliorer l'explicabilité et la détection des biais algorithmiques. Parmi celles-ci, l'ingénierie inversée est une approche consistant à analyser les entrées et les sorties d'un algorithme afin de comprendre comment il a pris ses décisions. Cependant, ce processus est complexe, chronophage et peu intuitif pour les personnes qui ne sont pas familières avec les techniques d'apprentissage automatique. Une approche plus accessible consiste à utiliser l'expertise des personnes compétentes avec la situation dans laquelle intervient un algorithme pour générer des contrefactuels pertinents. Un contrefactuel est observable lorsque la modification d'une variable d'entrée

entraîne un changement dans la sortie. Par exemple, si une modification du revenu ou de l'origine ethnique entraîne un résultat différent, cela peut indiquer un biais lié à une classification erronée ou révéler des déterminants ou des biais sociaux plus profonds qui doivent être pris en compte pour réduire les préjugés reproduits par l'algorithme (Panch et collab., 2019).

En matière d'explicabilité, nous sommes cependant aux prises avec un paradoxe lorsqu'on cherche à comprendre le fonctionnement des systèmes d'apprentissage automatique les plus complexes, tels que les réseaux de neurones. En théorie, comme nous venons de le présenter, plus un modèle est transparent, plus il devrait être facile de comprendre son fonctionnement et de repérer des erreurs et des risques de biais. Dans les systèmes d'apprentissage automatique complexes, une transparence accrue peut au contraire compliquer la compréhension, plutôt que la faciliter. Lorsque les modèles sont très sophistiqués, comme les réseaux de neurones profonds ou les modèles de renforcement, ils génèrent une grande quantité de détails techniques. Si toutes ces informations sont présentées de manière exhaustive aux utilisateurs, cela peut rapidement les submerger, surtout s'ils ne sont pas experts en la matière. Cette surcharge cognitive rend difficile la capacité de tirer des conclusions utiles, car l'abondance de détails masque les aspects essentiels, rendant l'interprétation et la détection des erreurs et des biais plus difficiles, voire impossibles (Cobbe, 2019).

Des auteurs soulignent que, dans des domaines sensibles, comme la justice pénale et la santé, où les décisions algorithmiques peuvent avoir des répercussions majeures sur la vie des personnes, la transparence est essentielle pour permettre une évaluation critique des décisions et établir un consensus sur la présence potentielle ou réelle de biais

(Goodman et Flaxman, 2017 ; Lloyd et Hamilton, 2018). Étant donné que les prédictions ou les décisions générées par des systèmes algorithmiques dans ces domaines sensibles peuvent avoir des conséquences directes et significatives sur les personnes concernées, les chercheurs plaident pour une exigence accrue en matière d'explicabilité. De plus, ces chercheurs recommandent de traiter l'utilisation de l'IA dans ces domaines sensibles, comme des questions de sécurité, en raison des risques liés aux erreurs ou aux biais algorithmiques (Henderson et collab., 2022). Cela signifie que les gouvernements devraient imposer une transparence complète aux développeurs d'algorithmes dans ces domaines sensibles et leur imposer l'obligation de donner un accès et une information sur les données d'entraînement, les méthodes et les techniques utilisées pour concevoir et entraîner l'algorithme et les objectifs de l'algorithme, c'est-à-dire ce qu'il cherche à atteindre, afin de s'assurer que ses actions sont en adéquation avec les attentes. De plus, des audits indépendants et périodiques pour évaluer la performance et la fiabilité des systèmes devraient être mis en place. En réalisant ces audits réguliers, des experts indépendants pourraient vérifier que le système d'IA reste performant à long terme et qu'il ne développe pas de biais ou de problèmes non détectés précédemment.

L'explicabilité est donc un facteur clé pour s'assurer que les systèmes d'IA génèrent des résultats fiables et prennent des décisions adéquates. Toutefois, l'explicabilité des systèmes algorithmiques ne se limite pas à une simple compréhension technique, elle englobe également une dimension humaine et sociale. Cela signifie que les décisions prises par des algorithmes doivent être compréhensibles non seulement par les experts techniques, mais aussi par les personnes concernées (utilisateurs, citoyens, usagers, décideurs, etc.). L'explicabilité prend en compte la capacité des individus à

saisir le raisonnement derrière une prédiction ou une décision (Besse et collab., 2021). En ce sens, l'explicabilité est une condition préalable pour que les organisations publiques puissent rendre des comptes, voire être tenues responsables des erreurs et des préjudices engendrés par un algorithme. Il est donc essentiel que les responsables politiques et les agents publics soient conscients des risques de discrimination liés à ces systèmes et qu'ils comprennent le fonctionnement des algorithmes afin de pouvoir se concentrer sur les aspects à surveiller. Un détour par le garage nous permettra d'illustrer la pertinence de cette recommandation. Si un automobiliste est conscient du risque de baisse de performance de son moteur et sait comment vérifier le niveau d'huile, il pourra surveiller lui-même l'état de son véhicule. En ne dépendant pas uniquement de son garagiste, il pourra intervenir rapidement et éviter ainsi des réparations coûteuses ou des dommages graves à son moteur. Améliorer la littératie numérique de l'ensemble de la population, en particulier des acteurs publics, souvent en première ligne de l'utilisation de ces systèmes et de leurs conséquences, permettrait d'acquérir une vigilance indispensable, notamment lors du déploiement des systèmes d'IA. Des efforts en matière de formation doivent être faits avec notamment la création de programmes spécifiques centrés sur l'utilisation de l'IA dans le secteur public. « Ces formations non seulement favorisent une culture de l'innovation, mais fournissent également aux employés du secteur public les moyens de tirer avantage de l'IA de manière éthique et efficace dans leurs activités quotidiennes, ce qui permet en fin de compte d'améliorer la fourniture de services et la réactivité [des organisations publiques] » (DEWG, 2024, p. 30, trad.).

* * *

Dans cette section, nous avons formulé les questions essentielles à se poser et les actions à entreprendre pour définir, atténuer ou éliminer les biais algorithmiques dans le secteur public. Pour approfondir cette analyse, il est possible de se référer à la grille de réflexion sur l'utilisation des algorithmes dans le secteur public, présentée en annexe.

En conclusion, il est important de rappeler que la recherche de solutions doit tenir compte du rythme soutenu de l'évolution technologique. En effet, cette évolution rapide permettrait de réduire, voire de résoudre complètement les problèmes associés aux biais algorithmiques, mais elle pourrait aussi changer la nature ou accélérer la production de biais.

À cet égard, il est essentiel d'adopter une approche proactive face à l'évolution rapide de la technologie, tout en reconnaissant que cette démarche est fondamentalement politique et doit s'inscrire dans le cadre du contrôle démocratique (Kuziemski et Misuraca, 2020). Il est nécessaire de réfléchir aux valeurs à promouvoir et aux visions du monde envisagées lors de l'utilisation des systèmes de décision automatisée dans le secteur public.

Six constats majeurs ressortent de la littérature existante sur l'utilisation d'algorithmes d'aide à la décision. Ces constats permettent de synthétiser les principaux enjeux présentés dans cet ouvrage :

- Il existe une forte incitation à utiliser les systèmes d'IA pour remplacer des agents publics jugés plus coûteux.
- Les systèmes de prise de décision basés sur l'IA peinent à être objectifs lorsqu'ils sont font face à des problèmes sociaux complexes.

- Des valeurs et des préférences doivent être intégrées dans le code de nombreux systèmes, ce qui peut conduire, consciemment ou non, à l'incorporation de biais et de préjugés humains.
- L'utilisation de systèmes de prise de décision automatisée a entraîné des conséquences négatives, parfois extrêmement dommageables, pour certaines personnes ou communautés, notamment les plus vulnérables.
- Les mécanismes de rétroaction [*feedback*] sont souvent absents ou incomplets, ce qui empêche les systèmes algorithmiques d'apprendre correctement ou les amène à tirer des conclusions erronées en se basant sur des données historiques parfois non pertinentes pour la situation actuelle.
- En plus des questions de justice et d'équité, les systèmes d'IA soulèvent des questions en matière de transparence, d'explicabilité et de responsabilité, notamment lorsque le développement et la mise à jour de ces systèmes sont externalisés à des entreprises du secteur privé[5] (Bannister et Connolly, 2020, p. 478, trad.).

5. Nous avons modifié l'ordre d'apparition des constats énoncés par Bannister et Connolly.

CONCLUSION

L'adoption croissante de l'IA par les organisations publiques soulève des questions importantes et complexes liées à l'équité, à la justice et à la responsabilité dans les processus décisionnels. Comme nous l'avons vu tout au long de cet ouvrage, les promesses d'efficacité offertes par l'IA se heurtent régulièrement à la réalité du problème des biais algorithmiques, capables de perpétuer, voire d'aggraver, les inégalités déjà présentes dans la société. En effet, les technologies, en particulier les algorithmes d'apprentissage automatique, offrent un potentiel dans de nombreux domaines d'action publique et peuvent contribuer à résoudre des problèmes complexes. Toutefois, les controverses emblématiques que nous avons recensées dans l'encadré 3 nous rappellent que l'intégration de l'IA dans les processus décisionnels nécessite une utilisation prudente et réfléchie. Cet avertissement est encore plus important lorsque les systèmes algorithmiques génèrent des prédictions ou alimentent des décisions qui affectent directement la vie des personnes. Comme nous l'avons vu, il s'agit d'un domaine dans lequel il reste du chemin à parcourir pour atteindre un niveau satisfaisant de qualité et de fiabilité (Waller et Waller, 2020).

Notre analyse met en évidence le fait que l'équité et la justice sont des caractéristiques des systèmes sociaux, plutôt

que des propriétés intrinsèques aux outils technologiques. L'ambition de parvenir à une administration numérique sans discrimination requiert une compréhension et une réflexion sur les inégalités qui existent dans notre société. Le phénomène résumé dans l'expression « mauvaises données à l'entrée, résultats erronés à la sortie » (*garbage in, garbage out*) nous rappelle que les algorithmes ne font que reproduire et, potentiellement, aggraver les biais sociaux existants. Cela met en évidence l'importance de la qualité et de la représentativité des données utilisées, ainsi que la nécessité d'une réflexion approfondie sur les objectifs et les valeurs que nous souhaitons intégrer dans les systèmes décisionnels automatisés. Le problème des biais algorithmiques dans l'administration publique constitue une occasion unique d'apprentissage. Tout comme l'IA nous amène à réfléchir à la nature de la pensée humaine autant qu'à celle de la machine, notre quête pour réduire les biais algorithmiques nous éclaire autant sur nos biais individuels, organisationnels ou sociétaux que sur les biais de l'algorithme. La véritable valeur de cette démarche réside dans cette réflexion collective et cette remise en question continue sur les biais et la quête d'équité dans la société et dans l'univers numérique.

Le principal défi que doivent relever les organisations publiques souhaitant utiliser l'IA réside dans la compréhension, l'identification, l'atténuation ou la suppression des biais d'une manière réfléchie et responsable. Des spécialistes nous mettent en garde sur les embûches qui risquent de surgir lorsque des efforts sont entrepris pour surmonter ce défi, car « la multitude des solutions proposées [pour atténuer ou éliminer les biais algorithmiques] est un reflet de la complexité du problème […] ; la recherche est en cours mais elle n'est pas parvenue à son terme. Du temps et des expérimentations seront encore nécessaires pour faire émerger les meilleures

solutions en fonction du contexte et des [technologies] » (Besse et collab., 2021, p. 24).

Nous sommes bien conscients que la transition vers une gouvernance de plus en plus algorithmique comporte d'autres défis importants. Elle exige de parvenir à un équilibre entre l'efficacité technique et l'équité sociale, tout en soulevant des questions philosophiques complexes liées à la justice distributive et au bien-être collectif. Les compromis entre la précision des prédictions et l'équité des résultats invitent à repenser nos conceptions de la justice sociale à l'ère numérique. Devant ces défis, il pourrait être tentant de renoncer à l'usage de l'IA dans le secteur public. Cependant, compte tenu de la rapidité des avancées technologiques et de la présence croissante de ces outils, une telle démarche paraît peu vraisemblable. Il est donc nécessaire de concentrer les efforts sur le développement de moyens visant à prévenir, réduire et éliminer les biais, tout en restant attentifs aux limites et aux risques propres aux technologies d'IA.

Depuis quelques années, les réflexions sur les enjeux et les défis relatifs à l'utilisation de l'IA dans le secteur public occupent une place grandissante dans les travaux du domaine de l'IA. Malgré cela, une minorité de pays se sont dotés de stratégies ou de réglementations spécifiques à l'utilisation de l'IA dans le secteur public. L'absence de cadre propre au secteur public est perçue par plusieurs spécialistes comme une lacune qui pourrait avoir des conséquences fâcheuses.

> Sans directives claires, les pratiques peuvent varier d'une organisation à l'autre, créant des disparités dans l'intégration et la régulation des technologies d'IA. Cette incohérence risque de nuire à la confiance du public envers les initiatives d'IA portées par les autorités publiques. De plus, des systèmes d'IA mal régulés peuvent accentuer les biais et les inégalités […]. Enfin, un manque de transparence dans les décisions automatisées compromet la responsabilité et la reddition de comptes (DEWG, 2024, p. 45, trad.).

L'intégration de l'IA dans le secteur public offre une occasion unique de repenser les processus décisionnels et de revisiter les principes fondamentaux de l'équité. En incitant les organisations à clarifier leurs objectifs et à formaliser leurs critères de décision, l'IA peut, paradoxalement, rendre les processus administratifs plus transparents et stimuler un débat essentiel sur les considérations éthiques des politiques et des interventions publiques. Plus largement, l'IA pourrait restructurer les opérations quotidiennes des administrations, en modifiant l'organisation du travail, la gestion ou l'allocation des ressources et la manière dont les usagers interagissent avec les organisations publiques. L'IA pourrait aussi, plus largement, transformer la culture administrative en introduisant de nouvelles méthodes de prise de décision basées sur des données et des algorithmes, changeant ainsi l'environnement global dans lequel les organisations publiques évoluent (Kuziemski et Misuraca, 2020 ; Paul et collab., 2024).

L'automatisation par l'IA ne réduit pas la responsabilité des agents et des organisations publiques ; au contraire, elle l'amplifie. Il est essentiel que les responsables politiques, les agents publics, les développeurs et la population collaborent pour s'assurer que les valeurs démocratiques, l'équité et la dignité soient au cœur des systèmes algorithmiques. Cela nécessite non seulement des solutions techniques, mais aussi des cadres réglementaires appropriés, une gouvernance transparente et un dialogue constant avec la société civile. La mise en place d'un cadre juridique approprié semble une nécessité, bien qu'elle présente certains défis (OGP, 2021). Les questions liées à la formation des décideurs et des agents publics, à la mise à jour et à l'amélioration des algorithmes ainsi qu'à la sensibilisation et à la participation de la communauté dans leur utilisation sont complexes et

méritent une attention particulière. Le fait que ces changements technologiques précèdent souvent l'établissement de cadres juridiques adaptés rend la tâche encore plus ardue et accentue l'urgence d'agir.

L'avenir de l'IA dans le secteur public reste à écrire, mais il est essentiel de le façonner de manière à ce qu'il incarne les valeurs d'équité et de justice. Relever ce défi offre l'occasion non seulement d'améliorer les services publics, mais aussi de renouveler les bases du contrat social qui unit l'État et la population.

ANNEXE

Grille de réflexion sur l'utilisation des algorithmes dans le secteur public

Dans cet ouvrage, nous avons présenté les biais algorithmiques dans le secteur public, en identifiant leurs sources potentielles et en proposant des solutions pour les prévenir, les atténuer ou les supprimer. Notre analyse met en lumière un point central : les biais algorithmiques ne sont pas uniquement des problèmes techniques et ils requièrent une intervention humaine pour être évités ou corrigés.

Afin d'accompagner les organisations publiques dans l'intégration et l'utilisation de l'IA tout en limitant les risques de biais, nous avons conçu une grille de réflexion regroupant des questions que les responsables politiques et les agents publics devraient se poser lorsqu'ils envisagent ou s'engagent à intégrer des algorithmes d'aide à la décision automatisée dans le secteur public.

1. DIMENSIONS TECHNIQUES	
1.1 Transparence et explicabilité	• L'architecture de l'algorithme est-elle documentée ? • Les sources de données et leurs méthodes de traitement sont-elles clairement expliquées ?
1.2 Qualité des données	• L'origine des données est-elle documentée et traçable ? • Les données sont-elles collectées de manière adéquate et respectent-elles les réglementations en vigueur ? • Les données sont-elles fiables, complètes et précises ? • Les données sont-elles à jour et reflètent-elles correctement les caractéristiques de la population ou les dimensions du problème ciblé ? • Une vérification a-t-elle été effectuée pour identifier des biais potentiels dans les données (biais démographiques, biais de sélection, etc.) ?
1.3 Sécurité des données	• Les protocoles de sécurité sont-ils en place pour protéger les données contre les cyberattaques et les atteintes à la confidentialité ? • Les données personnelles sont-elles anonymisées ou dépersonnalisées conformément aux réglementations ?
1.4 Fiabilité et performance	• L'algorithme a-t-il été testé pour détecter des biais dans les résultats ? • L'algorithme a-t-il été testé dans des conditions réelles ? • L'algorithme a-t-il été audité par des experts indépendants pour garantir sa performance ?
1.5 Suivi et évaluation	• Existe-t-il un processus de mise à jour continu de l'algorithme en fonction des nouvelles données ou des résultats des audits ? • Les indicateurs de performance sont-ils régulièrement suivis et ajustés ? • L'algorithme est-il ajusté régulièrement pour corriger les biais détectés ? • L'algorithme peut-il être mis à jour facilement pour maintenir ou améliorer sa performance ?

2. DIMENSIONS HUMAINES	
2.1 Acceptabilité sociale	• Comment la population perçoit-elle l'utilisation de cet algorithme dans le secteur public ? • Des mécanismes sont-ils mis en place pour impliquer la population dans le développement et l'évaluation de l'algorithme ? • L'utilisation de l'algorithme respecte-t-elle les droits fondamentaux ?
2.2 Pertinence	• Quel est le problème spécifique que cet algorithme cherche à résoudre ? • A-t-on clairement défini les besoins avant le développement de l'algorithme ? • Des alternatives non technologiques ont-elles été envisagées pour résoudre le problème identifié ? • L'algorithme tient-il compte de la diversité et de la complexité des situations humaines ou repose-t-il sur des simplifications excessives ?
2.3 Compréhension et contestation	• Est-il possible de comprendre et d'interpréter les décisions de l'algorithme ? • Les explications fournies sont-elles accessibles à des personnes non expertes ? • Les usagers sont-ils informés de la manière dont l'algorithme influence les décisions administratives ? • Les usagers ont-ils la possibilité de donner leur avis ou de contester les décisions prises par l'algorithme ?
2.4 Responsabilité humaine	• Comment l'organisation garantit-elle que l'algorithme ne remplace pas le jugement humain, surtout dans des contextes complexes ? • Des mécanismes sont-ils en place pour garantir qu'un humain puisse intervenir ou annuler une décision algorithmique si nécessaire ? • Les agents publics ont-ils les compétences adéquates pour comprendre et superviser le système algorithmique ?

2. DIMENSIONS HUMAINES (*suite*)	
2.5 Évaluation des impacts	• L'impact global de l'algorithme sur la société a-t-il été évalué ? • L'algorithme contribue-t-il à réduire ou à exacerber les inégalités sociales ou économiques ? • Des mécanismes sont-ils en place pour recueillir les commentaires des usagers concernant l'algorithme et son efficacité ? • Des mécanismes sont-ils en place pour mesurer les répercussions à long terme sur les usagers, les agents publics, la population en général et les groupes vulnérables en particulier ?

BIBLIOGRAPHIE

AI HLEG (High-Level Expert Group on Artificial Intelligence) (2018). *A Definition of AI : Main Capabilities and Scientific Disciplines*. European Commission.

Aizenberg, E. et J. van den Hoven (2020). « Designing for human rights in AI ». *Big Data & Society*, 7(2). <https://doi.org/10.1177/2053951720949566>

Alon-Barkat, S. et M. Busuioc (2023). « Human–AI interactions in public sector decision making : "Automation bias" and "selective adherence" to algorithmic advice ». *Journal of Public Administration Research and Theory*, 33(1), 153-169. <https://doi.org/10.1093/jopart/muac007>

Amnesty International (2021, 25 octobre). « Pays-Bas. Scandale des allocations familiales : un avertissement qui montre l'urgence d'interdire les algorithmes racistes ». <https://www.amnesty.org/fr/latest/news/2021/10/xenophobic-machines-dutch-child-benefit-scandal>

Angwin, J., J. Larson, S. Mattu et L. Kirchner (2016, 23 mai). « Machine Bias ». *ProPublica*. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Antony, L. M. (2016). « Bias : Friend or foe ? Reflections on saulish skepticism », dans M. Brownstein et J. Saul (dir.), *Implicit Bias and Philosophy*, vol. 1 : Metaphysics and Epistemology (p. 157-190). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198713241.003.0007>

Arène, R. (2024, 26 avril). « L'IA du fisc confond piscines et places handicapées : un taux d'erreur de 30 % révélé ». *CNET France*. <https://www.cnetfrance.fr/news/lia-du-fisc-confond-piscines-et-places-handicapees-un-taux-derreur-de-30-revele-404124.htm>

Arrêté ministériel nº 2024-02 du ministre de la Cybersécurité et du Numérique en date du 27 juin 2024 (2024, 7 août). *Gazette officielle du Québec*, nº 32, 5442-5457.

Assouline, M., S. Gilad et P. Ben-Nun Bloom (2022). « Discrimination of minority welfare claimants in the real world : The effect of implicit prejudice ». *Journal of Public Administration Research and Theory*, 32(1), 75-96. <https://doi.org/10.1093/jopart/muab016>

Autin, J.-L. (2011). « La motivation des actes administratifs unilatéraux, entre tradition nationale et évolution des droits européens ». *Revue française d'administration publique*, 137-138(1), 85-99. <https://doi.org/10.3917/rfap.137.0085>

Babuta, A. et M. Oswald (2019). *Data Analytics and Algorithmic Bias in Policing*. Briefing paper. Royal United Services Institute for Defence and Security Studies.

Baker, R. S. et A. Hawn (2022). « Algorithmic Bias in Education ». *International Journal of Artificial Intelligence in Education*, 32(4), 1052-1092. <https://doi.org/10.1007/s40593-021-00285-9>

Bannister, F. et R. Connolly (2020). « Administration by algorithm : A risk management framework ». *Information Polity*, 25(4), 471-490. <https://doi.org/10.3233/IP-200249>

Barocas, S., M. Hardt et A. Narayanan (2022). *Fairness and Machine Learning*. The MIT Press.

Berry, F. S. (1997). « Explaining managerial acceptance of expert systems ». *Public Productivity & Management Review*, 20(3), 323-335. <https://doi.org/10.2307/3380981>

Bertolucci, M. (2024). « L'intelligence artificielle dans le secteur public : revue de la littérature et programme de recherche ». *Gestion et management public*, 12(3), 71-91. <https://doi.org/10.3917/gmp.123.0071>

Besse, P., C. Castets-Renard et A. Garivier (2017). *Loyauté des décisions algorithmiques*. Contribution au débat public initié par la CNIL : Éthique et numérique.

Besse, P., A. Besse-Patin et C. Castets-Renard (2021). *Implications juridiques et éthiques des algorithmes d'intelligence artificielle dans le domaine de la santé*. Statistique et société.

Brennan, T., W. Dieterich et B. Ehret (2009). « Evaluating the predictive validity of the compas risk and needs assessment system ». *Criminal Justice and Behavior*, 36(1), 21-40. <https://doi.org/10.1177/0093854808326545>

Bullock, J. B. (2019). « Artificial intelligence, discretion, and bureaucracy ». *The American Review of Public Administration*, 49(7), 751-761. <https://doi.org/10.1177/0275074019856123>

Burrell, J. (2016). « How the machine "thinks" : Understanding opacity in machine learning algorithms ». *Big Data & Society*, 3(1). <https://doi.org/10.1177/2053951715622512>

Busch, P. A. et H. Z. Henriksen (2018). « Digital discretion : A systematic literature review of ICT and street-level discretion ». *Information Polity*, 23(1), 3-28. <https://doi.org/10.3233/IP-170050>

Busuioc, M. (2021). « Accountable artificial intelligence : Holding algorithms to account ». *Public Administration Review*, 81(5), 825-836. <https://doi.org/10.1111/puar.13293>

Chouldechova, A. et A. Roth (2020). « A snapshot of the frontiers of fairness in machine learning ». *Communications of the ACM*, 63(5), 82-89. <https://doi.org/10.1145/3376898>

Ciociola, A. A., C. Kefalas, L. B. Cohen, P. Kulkarni, A. Buchman, C. Burke, T. Cain, J. Connor, E. D. Ehrenpreis, J. Fang, R. Fass, R. Karlstadt, D. Pambianco, J. Phillips, M. Pochapin, P. Pockros, P. Schoenfeld, R. Vuppalanchi et FDA-Related Matters Committee of the American College of Gastroenterology (2014). « How drugs are developed and approved by the FDA : Current process and future directions ». *The American Journal of Gastroenterology*, 109(5), 620-623. <https://doi.org/10.1038/ajg.2013.407>

Cluzel-Métayer, L. (2018). « Calculer : l'influence des algorithmes sur l'édiction des décisions administratives ». Dans Association française pour la recherche en droit administratif (AFDA) (dir.). *Les méthodes en droit administratif* (p. 243-259). Dalloz.

Cluzel-Métayer, L. (2020). « Décision publique algorithmique et discriminations ». Dans M. Mercat-Bruns (dir.). *Nouveaux modes de détection et de prévention de la discrimination et accès au droit : action de groupe et discrimination systémique : algorithmes et préjugés ;*

réseaux sociaux et harcèlement (p. 109-119). Société de législation comparée.

CNIL (Commission nationale de l'informatique et des libertés) (s.d.). « Algorithme ». <https://www.cnil.fr/fr/definition/algorithme>

CNIL (Commission nationale de l'informatique et des libertés) (2023, 11 octobre). « Intelligence artificielle : la CNIL dévoile ses premières réponses pour une IA innovante et respectueuse de la vie privée ». <https://www.cnil.fr/fr/intelligence-artificielle-la-cnil-devoile-ses-premieres-reponses-pour-une-ia-innovante-et>

Cobbe, J. (2019). « Administrative law and the machines of government : Judicial review of automated public-sector decision-making ». *Legal Studies*, 39(4), 636-655. <https://doi.org/10.1017/lst.2019.9>

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

Corbett-Davies, S., J. D. Gaebler, H. Nilforoshan, R. Shroff et S. Goel (2023). « The measure and mismeasure of fairness ». *Journal of Machine Learning Research*, 24. <https://www.jmlr.org/papers/volume24/22-1511/22-1511.pdf>

Corbett-Davies, S., E. Pierson, A. Feller, S. Goel et A. Huq (2017). « Algorithmic decision making and the cost of fairness ». *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17).* Association for Computing Machinery, New York, 797-806. <https://doi.org/10.1145/3097983.309809>

Cormen, T. H., C. E. Leiserson, R. L. Rivest et C. Stein (2022). *Introduction to Algorithms*. 4e éd., MIT Press.

Cossette-Lefebvre, H. et J. Maclure (2022). « AI's fairness problem : Understanding wrongful discrimination in the context of automated decision-making ». *AI and Ethics*, 3, 1255-1269. <https://doi.org/10.1007/s43681-022-00233-w>

CSF (Conseil du statut de la femme) (2023). « Intelligence artificielle : des risques pour l'égalité entre les femmes et les hommes ». Gouvernement du Québec. <https://csf.gouv.qc.ca/wp-content/uploads/Avis_intelligence_artificielle.pdf>

Dai, J., S. Fazelpour et Z. Lipton (2021). « Fair machine learning under partial compliance ». *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Association for Computing Machinery, New York, 55-65. <https://doi.org/10.1145/3461702.3462521>

Danks, D. et A. J. London (2017). « Algorithmic bias in autonomous systems ». *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 4691-4697. <https://doi.org/10.24963/ijcai.2017/654>

Dean, S., T. K. Gilbert, N. Lambert et T. Zick (2021). « Axes for sociotechnical inquiry in AI research ». *IEEE Transactions on Technology and Society*, 2(2), 62-70. <https://doi.org/10.1109/TTS.2021.3074097>

De-Arteaga, M., R. Fogliato et A. Chouldechova (2020). « A case for humans-in-the-loop : Decisions in the presence of erroneous algorithmic scores ». *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, 1-12. <https://doi.org/10.1145/3313831.3376638>

Demartini, G., K. Roitero et S. Mizzaro (2021). « Managing bias in human-annotated data : Moving beyond bias removal ». (arXiv :2110.13504). *arXiv*. <https://doi.org/10.48550/arXiv.2110.13504>

Demon, V. (2023, 16 avril). « VioGén, mesure-phare de l'Espagne contre les violences faites aux femmes ». *RTS*. <https://www.rts.ch/info/monde/13939642-viogen-mesurephare-de-lespagne-contre-les-violences-faites-aux-femmes.html>

Desouza, K. C., G. S. Dawson et D. Chenok (2020). « Designing, developing, and deploying artificial intelligence systems : Lessons from and for the public sector ». *Business Horizons*, 63(2), 205-213. <https://doi.org/10.1016/j.bushor.2019.11.004>

DEWG (Digital Economy Working Group) (2024). *Mapping the Development, Deployment and Adoption of AI for Enhanced Public Services in the G20 Members : Artificial Intelligence for Inclusive Sustainable Development and Inequalities Reduction*.

Diakopoulos, N. (2015). « Algorithmic Accountability : Journalistic investigation of computational power structures ». *Digital Journalism*, 3(3), 398-415. <https://doi.org/10.1080/21670811.2014.976411>

Dietvorst, B. J., J. P. Simmons et C. Massey (2015). « Algorithm aversion : People erroneously avoid algorithms after seeing them err ». *Journal of Experimental Psychology : General*, 144(1), 114-126. <https://doi.org/10.1037/xge0000033>

Dobbe, R., S. Dean, T. Gilbert et N. Kohli (2018). « A broader view on bias in automated decision-making : Reflecting on epistemology and dynamics » (arXiv :1807.00553). *arXiv*. <http://arxiv.org/abs/1807.00553>

Eckhouse, L., K. Lum, C. Conti-Cook et J. Ciccolini (2019). « Layers of bias : A unified approach for understanding problems with risk assessment ». *Criminal Justice and Behavior*, 46(2), 185-209. <https://doi.org/10.1177/0093854818811379>

ELA (European Labour Authority) (2023). *Artificial Intelligence and Algorithms in Risk Assessment : Addressing Bias, Discrimination and other Legal and Ethical Issues*. A Handbook. <https://www.ela.europa.eu/sites/default/files/2023-08/ELA-Handbook-AI-training.pdf>

Engstrom, D. F., D. E. Ho, C. M. Sharkey et M.-F. Cuéllar (2020). « Government by algorithm : Artificial intelligence in federal administrative agencies ». *Public Law Research Paper*, nº 20-54, New York University, School of Law. <http://dx.doi.org/10.2139/ssrn.3551505>

Ensign, D., S. A. Friedler, S. Neville, C. Scheidegger et S. Venkatasubramanian (2017). « Runaway feedback loops in predictive policing » (arXiv :1706.09847). *arXiv*. <http://arxiv.org/abs/1706.09847>

Eticas.ai (2023). « The adversarial audit of VioGén : Three years later and new system version ». <https://eticas.ai/the-adversarial-audit-of-viogen-three-years-later/>

Eubanks, V. (2018). *Automating Inequality : How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, Inc.

Fazelpour, S. et D. Danks (2021). « Algorithmic bias : Senses, sources, solutions ». *Philosophy Compass*, 16(8), e12760. <https://doi.org/10.1111/phc3.12760>

Ferguson, A. G. (2017). *The Rise of Big Data Policing : Surveillance, Race, and the Future of Law Enforcement*. New York University Press.

Franzen, S., C. Quang, L. Schweizer, A. Budzier, J. Gold, M. Vellez, S. Ramirez et E. Raimondo (2022). *Advanced Content Analysis : Can Artificial Intelligence Accelerate Theory-driven Complex Program Evaluation ?* Independent Evaluation Group, World Bank Group.

Friedler, S. A., C. Scheidegger et S. Venkatasubramanian (2016). « On the (im)possibility of fairness » (arXiv :1609.07236). *arXiv*. <http://arxiv.org/abs/1609.07236>

Friedman, B. et H. Nissenbaum (1996). « Bias in computer systems ». *ACM Transactions on Information Systems*, 14(3), 330-347. <https://doi.org/10.1145/230538.230561>

Gaon, A. et I. Stedman (2019). « A call to action : Moving forward with the governance of artificial intelligence in Canada ». *Alberta Law Review*, 56(4), 1137-1165. <https://doi.org/10.29173/alr2547>

Geiger, G., S. Pénicaud, M. Romain et A. Sénécat (2023, 4 décembre). « Profilage et discriminations : enquête sur les dérives des caisses d'allocations familiales ». *Le Monde*.

Gerke, S., B. Babic, T. Evgeniou et I. Cohen (2020). « The need for a system view to regulate artificial intelligence/machine learning-based software as medical device ». *NPJ Digital Medicine*, 3. <https://doi.org/10.1038/s41746-020-0262-2>

Goodman, B. et S. Flaxman (2017). « European Union regulations on algorithmic decision making and a "right to explanation" ». *AI Magazine*, 38(3), 50-57. <https://doi.org/10.1609/aimag.v38i3.2741>

Gouvernement du Québec (2021). *Stratégie d'intégration de l'intelligence artificielle dans l'administration publique 2021-2026*. <https://www.quebec.ca/gouvernement/politiques-orientations/vitrine-numeriqc/strategie-integration-ia-administration-publique-2021-2026/mettre-a-profit-les-avancees-de-ia-au-service-du-secteur-public>

GPAI (Global Partnership on Artificial Intelligence) (2024). *Algorithmic Transparency in the Public Sector : A State-of-the-art Report of Algorithmic Transparency Instruments*. <https://gpai.ai/projects/responsible-ai/algorithmic-transparency-in-the-public-sector/algorithmic-transparency-in-the-public-sector.pdf>

Green, B. et Y. Chen (2019). « Disparate interactions : An algorithm-in-the-loop analysis of fairness in risk assessments ». *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 90-99. <https://doi.org/10.1145/3287560.3287563>

Gribonval, R. (2021, 2 février). « D'APB à Parcoursup : quelles méthodes d'affectation post-bac ? » *Interstices*. <https://interstices.info/dapb-a-parcoursup-quelles-methodes-daffectation-post-bac/>

Harcourt, B. E. (2018). « The systems fallacy : A genealogy and critique of public policy and cost-benefit analysis ». *Journal of Legal Studies*, 47(2), 419-447. <https://chicagounbound.uchicago.edu/jls/vol47/iss2/7>

Hellström, T., V. Dignum et S. Bensch (2020). « Bias in machine learning what is it good (and bad) for ? » (arXiv : 2004.00686). *arXiv*. <https://doi.org/10.48550/arXiv.2004.00686>.

Henderson, B., C. M. Flood et T. Scassa (2022). « Artificial intelligence in Canadian healthcare : Will the law protect us from algorithmic bias resulting in discrimination ? » *Canadian Journal of Law and Technology*, 19(2), 475-504.

Hupe, P. et M. Hill (2007). « Street-level bureaucracy and public accountabiliy ». *Public Administration*, 85(2), 279-299. <https://doi.org/10.1111/j.1467-9299.2007.00650.x>

IEEE Standards Association (2023, 19 juillet). « The bias against bias : Using bias intentionally in artificial intelligence ». <https://standards.ieee.org/beyond-standards/the-bias-against-bias/>

Jacob, S. (2024). « Navigating the challenges of policy evaluation ». *Canadian Public Administration*, 67(2), 282-290. <https://doi.org/10.1111/capa.12571>

Jacob, S. (2025). « Artificial intelligence and the future of evaluation : From augmented to automated evaluation ». *Digital Government : Research and Practice*, 6(1), 1-10. <https://dl.acm.org/doi/10.1145/3696009>

Jacob, S. et S. Brousseau (2024). *L'IA dans le secteur public : cas d'utilisation et enjeux éthiques*. Obvia.

Jacob, S. et J. Lawarée, (2022). *Les mesures publiques dans les stratégies gouvernementales en matière d'intelligence artificielle : une perspective internationale*. Obvia.

Jacob, S. et S. Souissi, (2022). « Les intervenants sociaux face à la transformation numérique : synthèse de littérature internationale sur l'évolution de la mission et des compétences professionnelles ». *Revue des politiques sociales et familiales*, (145), 83-93. <https://doi.org/10.3917/rpsf.145.0083>

Jilke, S. et L. Tummers (2018). « Which clients are deserving of help ? A theoretical model and experimental test ». *Journal of Public Administration Research and Theory*, 28(2), 226-238. <https://doi.org/10.1093/jopart/muy002>

Johnson, G. M. (2021). « Algorithmic bias : On the implicit biases of social technology ». *Synthese,* 198(10), 9941-9961. <https://doi.org/10.1007/s11229-020-02696-y>

Jones, N. (2015, 29 juillet). « Anne Tolley scraps "lab rat" study on children ». *The New Zealand Herald*. <https://www.nzherald.co.nz/nz/anne-tolley-scraps-lab-rat-study-on-children/C7GIGYW2467HG327FKXFRJDPEM/>

Katzenbach, C. et L. Ulbricht (2019). « Algorithmic governance ». *Internet Policy Review*, 8(4). <https://doi.org/10.14763/2019.4.1424>

Kearns, M. et A. Roth, A. (2020). *The Ethical Algorithm : The Science of Socially Aware Algorithm Design*. Oxford University Press.

Kemper, L., G. Vorhoff et B. U. Wigger (2020). « Predicting student dropout : A machine learning approach ». *European Journal of Higher Education*, 10(1), 28-47. <https://doi.org/10.1080/21568235.2020.1718520>

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig et S. Mullainathan (2017). « Human decisions and machine predictions ». *The Quarterly Journal of Economics*, 133(1), 237-293. <https://doi.org/10.1093/qje/qjx032>

Kleinberg, J., J. Ludwig, S. Mullainathan et C. R. Sunstein (2018). « Discrimination in the age of algorithms ». *Journal of Legal Analysis*, 10, 113-174. <https://doi.org/10.1093/jla/laz001>

Kolkman, D. (2020, 26 août). « "F**k the algorithm"? What the world can learn from the UK's A-level grading fiasco ». *LSE Impact Blog*. <https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/>

König, P. D. et G. Wenzelburger (2021). « The legitimacy gap of algorithmic decision-making in the public sector : Why it arises and how to address it ». *Technology in Society*, 67, 101688. <https://doi.org/10.1016/j.techsoc.2021.101688>

Kuziemski, M. et G. Misuraca (2020). « AI governance in the public sector : Three tales from the frontiers of automated decision-making in democratic settings ». *Telecommunications Policy*, 44(6), 101976. <https://doi.org/10.1016/j.telpol.2020.101976>

Larson, J., S. Mattu, L. Kirchner et J. Angwin (2016, 23 mai). « How we analyzed the COMPAS recidivism algorithm ». *ProPublica*. <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>

Lattimore, F., S. O'Callaghan, Z. Paleologos, A. Reid, E. Santow, H. Sargeant et A. Thomsen (2020). *Using Artificial Intelligence to Make Decisions : Addressing the Problem of Algorithmic Bias*. Technical paper. Australian Human Rights Commission.

Le Galès, P. (2019). « Gouvernance ». Dans L. Boussaguet, S. Jacquot et P. Ravimet (dir.), *Dictionnaire des politiques publiques*. 5e édition (p. 297-305). Presses de Sciences po.

Levy, K., K. E. Chasalow et S. Riley (2021). « Algorithms and decision-making in the public sector ». *Annual Review of Law and Social Science*, 17(1), 309-334. <https://doi.org/10.1146/annurev-lawsocsci-041221-023808>

Lloyd, K. et B. A. Hamilton (2018). « Bias amplification in artificial intelligence systems » (arXiv : 1809.07842). *arXiv*. <https://doi.org/10.48550/arXiv.1809.07842>

Lum, K. et W. Isaac (2016). « To predict and serve ? » *Significance*, 13(5), 14-19. <https://doi.org/10.1111/j.1740-9713.2016.00960.x>

Mao, F. (2023, 7 juillet). « Robodebt : Illegal Australian welfare hunt drove people to despair ». *BBC News*. <https://www.bbc.com/news/world-australia-66130105>

Marique, E. et A. Strowel (2017). « Gouverner par la loi ou les algorithmes : de la norme générale de comportement au guidage rapproché des conduites ». *Dalloz IP/IT*, nº 10, 517-521. <https://dial.uclouvain.be/pr/boreal/object/boreal:188689>

Mayson, S. G. (2019). « Bias in, bias out ». *The Yale Law Journal*, 2218-2300. <https://www.yalelawjournal.org/article/bias-in-bias-out>

MCN (Ministère de la Cybersécurité et du Numérique) (2024a). « Potentialités de l'IA dans l'administration publique ». Gouvernement du Québec. <https://www.quebec.ca/gouvernement/politiques-orientations/vitrine-numeriqc/strategie-integration-ia-administration-publique-2021-2026/potentialites-ia-administration-publique>

MCN (Ministère de la Cybersécurité et du Numérique) (2024b). « Principes de mise en œuvre de la Stratégie d'intégration de l'intelligence artificielle dans l'administration publique ». Gouvernement du Québec. <https://www.quebec.ca/gouvernement/politiques-orientations/vitrine-numeriqc/strategie-integration-ia-administration-publique-2021-2026/principes-mise-oeuvre-strategie-ia>

McQuillan (2019, 7 juin). « AI realism and structural alternatives ». <https://www.danmcquillan.org/ai_realism.html>

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman et A. Galstyan (2022). « A survey on bias and fairness in machine learning ». *ACM Computing Surveys*, 54(6), 1-35. <https://doi.org/10.1145/3457607>

Miller, A. P. (2018, 26 juillet). « Want less-biased decisions ? Use algorithms ». *Harvard Business Review*. <https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms>

Milli, S., J. Miller, A. D. Dragan et M. Hardt (2018). « The social cost of strategic classification ». (arXiv:1808.08460). *arXiv*. <https://doi.org/10.48550/arXiv.1808.08460>

Mitchell, S., E. Potash, S. Barocas, A. D'Amour et K. Lum (2021). « Algorithmic fairness : Choices, assumptions, and definitions ». *Annual Review of Statistics and Its Application*, 8(1), 141-163. <https://doi.org/10.1146/annurev-statistics-042720-125902>

Montagnon, E., M. Cerny, A. Cadrin-Chênevert et collab. (2020) « Deep learning workflow in radiology : A primer ». *Insights Imaging* 11, 22. <https://doi.org/10.1186/s13244-019-0832-5>

Musiani, F. (2013). « Governance by algorithms ». *Internet Policy Review*, 2(3). <https://doi.org/10.14763/2013.3.188>

Newman, J. et M. Mintrom (2023). « Mapping the discourse on evidence-based policy, artificial intelligence, and the ethical practice of policy analysis ». *Journal of European Public Policy*, 30(9), 1839-1859. <https://doi.org/10.1080/13501763.2023.2193223>

Nkonde, M. (2023, 27 février). « How ChatGPT could perpetuate AI racism and bias. Here's how this powerful tech can become more accurate and inclusive ». *Mashable*. <https://mashable.com/article/chatgpt-ai-racism-bias>

Obermeyer, Z., B. Powers, C. Vogeli et S. Mullainathan (2019). « Dissecting racial bias in an algorithm used to manage the health of populations ». *Science*, 366(6464), 447-453. <https://www.science.org/doi/10.1126/science.aax2342>

Ocumpaugh, J., R. Baker, S. Gowda, N. Heffernan et C. Heffernan (2014). « Population validity for educational data mining models : A case study in affect detection ». *British Journal of Educational Technology*, 45(3), 487-501. <https://doi.org/10.1111/bjet.12156>

OECD (Organisation for Economic Co-operation and Development) (2024), « Governing with artificial intelligence : Are governments ready ? » *OECD Artificial Intelligence Papers*, nº 20, OECD Publishing. <https://doi.org/10.1787/26324bc2-en>

OECD.AI (2021). « Database of national AI policies ». <https://oecd.ai>

OGP (Open Government Partnership) (2021, 19 septembre). « Garantir la confiance dans la prise de décision publique qui utilise des algorithmes ». *Open Stories*. <https://www.ogpstories.org/fr/ensuring-confidence-in-public-decision-making-that-uses-algorithms/>

O'Neil, C. (2016). *Weapons of Math Destruction : How Big Data Increases Inequality and Threatens Democracy*. Penguin Random House.

Oswald, M. (2018). « Algorithm-assisted decision-making in the public sector : Framing the issues using administrative law rules governing discretionary power ». *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences*, 376(2128), 20170359. <https://doi.org/10.1098/rsta.2017.0359>

Paige, M. et A. Amrein-Beardsley (2020). « "Houston, we have a lawsuit" : A cautionary tale for the implementation of value-added models for high-stakes employment decisions ». *Educational Researcher*, 49(5), 350-359. <https://doi.org/10.3102/0013189X20923046>

Panch, T., H. Mattie et R. Atun (2019). « Artificial intelligence and algorithmic bias : Implications for health systems ». *Journal of Global Health*, 9(2), 010318. <https://doi.org/10.7189/jogh.09.020318>

Paquette, L., A. Andres, J. Ocumpaugh, R. Baker et Z. Li (2020). « Who's learning ? Using demographics in EDM research ». *Journal of Educational Data Mining*, 12(3), 1-30. <https://doi.org/10.5281/zenodo.4143612>

Paul, R., E. Carmel et J. Cobbe (dir.) (2024). *Handbook on Public Policy and Artificial Intelligence*. Edward Elgar Publishing Limited.

Pedersen, M. J., J. M. Stritch et F. Thuesen (2018). « Punishment on the frontlines of public service delivery : Client ethnicity and caseworker sanctioning decisions in a Scandinavian welfare state ». *Journal of Public Administration Research and Theory*, 28(3), 339-354. <https://doi.org/10.1093/jopart/muy018>

Petrogradskaya, A.A., A. P. Korobova et V. K. Barchukov (2021). « Intellectual management of the budget process in municipalities ». Dans S. I. Ashmarina, J. Horák, J. Vrbka et P. Šuleř (dir.) *Economic Systems in the New Era : Stable Systems in an Unstable World. IES 2020*. Lecture Notes in Networks and Systems, 160 (p. 713-718). Springer. <https://doi.org/10.1007/978-3-030-60929-0_92>

Poussart, D. (2021). « Gouvernance algorithmique, intelligence artificielle, enjeux, éthique : esquisse d'une analyse critique ». *Éthique publique*, 23(2). <https://doi.org/10.4000/ethiquepublique.6418>

Powles, J. et H. Nissenbaum (2018, 7 décembre). « The seductive diversion of "solving" bias in artificial intelligence ». *Medium*. <https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53>

Pūraitė, A., V. Zuzevičiūtė, D. Bereikienė, T. Skrypko et L. Shmorgun (2020). « Algorithmic governance in public sector : Is digitization a key to effective management ». *Independent Journal of Management & Production*, 11(9), 2149-2170. <https://doi.org/10.14807/ijmp.v11i9.1400>

Raaphorst, N. (2021). « Administrative justice in street-level decision-making : Equal treatment and responsiveness ». Dans M. Hertogh, R. Kirkham, R. Thomas et J. Tomlinson (dir.), *The Oxford Handbook of Administrative Justice* (p. 397-414). Oxford Handbooks.

Rambachan, A. et J. Roth (2020). « Bias in, bias out ? Evaluating the folk wisdom ». Dans 1st Symposium on Foundations of Responsible Computing (FORC 2020). *Leibniz International Proceedings in Informatics* (LIPIcs), 156, 6 :1-6 :15. <https://doi.org/10.4230/LIPIcs.FORC.2020.6>

Ramineni, C. et D. M. Williamson (2013). « Automated essay scoring : Psychometric guidelines and practices ». *Assessing Writing*, 18(1), 25-39. <https://doi.org/10.1016/j.asw.2012.10.004>

Rinta-Kahila, T., I. Someh, N. Gillespie, M. Indulska et S. Gregor (2024). « Managing unintended consequences of algorithmic decision-making : The case of Robodebt ». *Journal of Information Technology Teaching Cases*, 14(1), 165-171. <https://doi.org/10.1177/20438869231165538>

Ritter, S., M. Yudelson, S. E. Fancsali et S. R. Berman (2016). « How mastery learning works at scale ». *Proceedings of the Third (2016) ACM Conference on Learning @ Scale* (p. 71-79). <https://doi.org/10.1145/2876034.2876039>

Scott, P. G. (1997). « Assessing determinants of bureaucratic discretion : An experiment in street-level decision making ». *Journal of Public Administration Research and Theory*, 7(1), 35-58.

Selbst, A. D., D. Boyd, S. A. Friedler, S. Venkatasubramanian et J. Vertesi (2019). « Fairness and abstraction in sociotechnical systems ». *Proceedings of the Conference on Fairness, Accountability, and Transparency* (p. 59-68). <https://doi.org/10.1145/3287560.3287598>

Sénéchal, J.-F. (21 décembre 2023). *IA Café – Enquête au cœur de la recherche sur l'intelligence artificielle*, nº 76. Vidéo YouTube. <https://www.youtube.com/playlist?list=PL_LVlLa3F3JNgxC946lIk4c4k3j727E5I>

Silva, S. et M. Kenney (2019). « Algorithms, platforms, and ethnic bias ». *Communications of the ACM*, 62(11), 37-39. <https://doi.org/10.1145/3318157>

Srinivasan, R. et A. Chander (2021). « Biases in AI systems ». *Communications of the ACM*, 64(8), 44-49. <https://doi.org/10.1145/3464903>

Supiot, A. (2015). *La gouvernance par les nombres : cours au collège de France (2012-2014)*. Fayard.

Surden, H. (2020). « Ethics of AI in law : Basic questions ». Dans M. D. Dubber, F. Pasquale et S. Das (dir.), *The Oxford Handbook of Ethics of AI* (p.719-736). Oxford Academic.

Suresh, H. et J. V. Guttag (2021). « A framework for understanding sources of harm throughout the machine learning life cycle ». *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (p. 1-9). <https://doi.org/10.1145/3465416.3483305>

Susskind, J. (2018). *Future Politics : Living Together in a World Transformed by Tech*. Oxford University Press.

Torralba, A. et A. A. Efros (2011). « Unbiased look at dataset bias ». *CVPR 2011* (p. 1521-1528). IEEE Xplore.

Udoh, E. S. (2020). « Is the data fair ? : An assessment of the data quality of algorithmic policing systems ». *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance* (p. 1-7). <https://doi.org/10.1145/3428502.3428503>

USAID (United States Agency for International Development) (2022). *Managing Machine Learning Projects in International Development: A Practical Guide*. <https://www.usaid.gov/sites/default/files/2022-05/Vital_Wave_USAID-AIML-FieldGuide_FINAL_VERSION_1.pdf>

van Giffen, B., D. Herhausen et T. Fahse (2022). « Overcoming the pitfalls and perils of algorithms : A classification of machine learning biases and mitigation methods », *Journal of Business Research*, 144(C), 93-106. <https://doi.org/10.1016/j.jbusres.2022.01.076>

Van Voorhis, S. (2023, 7 mars). « ChatGPT : Did Big Tech set up the world for an AI bias disaster ? », *Harvard Business School Working Knowledge*. <https://www.library.hbs.edu/working-knowledge/chatgpt-did-big-tech-set-up-the-world-for-ai-bias-disaster>

Veale, M. et I. Brass (2019). « Administration by algorithm ? Public management meets public sector machine learning », Dans K. Yeung et M. Lodge (dir.). *Algorithmic Regulation* (p. 121-149). Oxford University Press. <https://doi.org/10.1093/oso/9780198838494.003.0006>

Villani, C. (2018). *Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne*. <https://www.vie-publique.fr/files/rapport/pdf/184000159.pdf>

Vogl, T. M., C. Seidelin, B. Ganesh et J. Bright (2020). « Smart technology and the emergence of algorithmic bureaucracy : Artificial intelligence in UK local authorities », *Public Administration Review*, 80(6), 946-961. <https://doi.org/10.1111/puar.13286>

Waller, M. et P. Waller (2020, 21 octobre). *Why Predictive Algorithms Are so Risky for Public Sector Bodies*. <https://doi.org/10.2139/ssrn.3716166>

Wang, D., S. Prabhat et N. Sambasivan (2022). « Whose AI dream ? In search of the aspiration in data annotation » *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (p. 1-16). <https://doi.org/10.1145/3491102.3502121>

Waters, A. et R. Miikkulainen (2014). « GRADE : Machine-learning support for graduate admissions ». *AI Magazine*, 35(1). <https://doi.org/10.1609/aimag.v35i1.2504>

Williams, A., M. Miceli et T. Gebru (2022, 13 octobre). « The exploited labor behind artificial intelligence », *Noema*. <https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/>

Wirtz, B. W. et W. M. Muller (2019). « An integrated artificial intelligence framework for public management », *Public Management Review*, 21(7), 1076-1100.

Yeung, K. (2018). « Algorithmic regulation : A critical interrogation », *Regulation & Governance*, 12(4), 505-523. <https://doi.org/10.1111/rego.12158>

Young, M. M., J. B. Bullock et J. D. Lecy (2019). « Artificial discretion as a tool of governance : A framework for understanding the impact of artificial intelligence on public administration », *Perspectives on Public Management and Governance*, 2(4), 301-313. <https://doi.org/10.1093/ppmgov/gvz014>

Young, M.M., J. Himmelreich, J. B. Bullock et K.-C. Kim (2021), « Artificial intelligence and administrative evil », *Perspectives on Public Management and Governance*, 4(3) 244-258. <https://doi.org/10.1093/ppmgov/gvab006>

Zuboff, S. (2020). *L'âge du capitalisme de surveillance : le combat pour un avenir humain face aux nouvelles frontières du pouvoir*. Zulma.

ÉTHIQUE, IA ET SOCIÉTÉ

Collection dirigée par Lyse Langlois

Autres titres parus dans la collection Éthique IA et société

Schallum Pierre et Fehmi Jaafar (dir.), *Médias sociaux : perspectives sur les défis liés à la cybersécurité, la gouvernementalité algorithmique et l'intelligence artificielle*, 2020.

Karine Gentelet (dir.). *Les intelligences artificielles au prisme de la justice sociale / Considering Artificial Intelligence Through the Lens of Social Justice*, 2023.

Félix Pageau, Tenzin Wagmo et Emilian Mihailov (dir.), *Robots and Gadgets : Aging at Home*, 2024.

Cécile Petitgand, *Données personnelles : reprenons le pouvoir ! Réflexions citoyennes sur la gouvernance à l'ère numérique*, 2024.

Véronique Guèvremont et Colette Brin (dir.), *Intelligence artificielle, culture et médias*, 2024.

L'introduction de l'intelligence artificielle (IA) dans les organisations publiques transforme radicalement la manière dont les décisions sont prises et soulève des inquiétudes en ce qui concerne la déshumanisation des services, les atteintes à la vie privée ou les risques de discrimination. Des décisions importantes sont maintenant prises à l'aide d'algorithmes dans de nombreux domaines d'intervention de l'État, que ce soit dans l'attribution d'aides sociales, la protection des enfants vulnérables, la détection des fraudes, la prédiction du décrochage scolaire, la lutte contre les violences faites aux femmes ou encore la gestion des libérations conditionnelles.

Les auteurs présentent des cas concrets d'utilisation de l'IA dans le secteur public, en se concentrant particulièrement sur les risques de biais algorithmiques dans la prise de décision. Ils proposent des solutions concrètes pour prévenir, corriger ou supprimer les biais algorithmiques. Cet ouvrage deviendra une lecture incontournable pour les responsables politiques, les membres de la fonction publique, les spécialistes du numérique et toutes les personnes intéressées par l'avenir des services publics.

STEVE JACOB est professeur titulaire de science politique à l'Université Laval et directeur scientifique à l'Observatoire international sur les impacts sociétaux de l'IA et du numérique (Obvia). Il mène des recherches sur la transformation numérique des organisations publiques, l'éthique publique et les dispositifs d'évaluation et de gestion de la performance.

SÉBASTIEN BROUSSEAU est sociologue de formation et professionnel de recherche à l'Obvia. Il mène des recherches interdisciplinaires sur les interactions entre l'humain et la technologie. Depuis 2018, il se spécialise en matière d'intelligence artificielle et de technologies émergentes en explorant leur dimension sociotechnique.

Illustration de couverture : iStockphoto

Série Points de vue

ÉTHIQUE, IA ET SOCIÉTÉ
Collection dirigée par Lyse Langlois

ISBN 978-2-7663-0724-1
9 782766 307241

Presses de l'Université Laval