RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

IGPDE

Intelligence Artificielle : Intelligence de l'Action publique ?

REVUE DE LITTÉRATURE

La revue de littérature ci-dessous a été réalisée par Edoardo Ferlazzo, chef du département de gestion publique comparée au sein du bureau de la Recherche de l'IGPDE[1]. Utilisée pour construire la trame du dessin animé « Intelligence artificielle : Intelligence de l'Action publique ? », elle s'appuie sur des travaux académiques afin de revenir sur l'histoire de l'intelligence artificielle, son développement, les concepts qu'elle porte, pour aider à mieux comprendre son objet et le type de questions que son déploiement et son usage impliquent, notamment au sein de l'action publique.

Depuis quelques années, tout particulièrement à la faveur du développement des méthodes d'apprentissage profond utilisant des réseaux de neurones, dites de *deep learning*, l'Intelligence artificielle (IA) n'en finit plus d'alimenter espoirs, craintes, interrogations et controverses au sein de l'espace public (Crépel et Cardon, 2022 ; Bellon et Velkowska, 2024 ; Vayre et Gaglio, 2020). Pour certains, l'IA place l'humanité à l'aube d'une quatrième révolution industrielle et de solutions technologiques de rupture automatisant voire améliorant les décisions humaines. Pour d'autres, l'IA présenterait des risques majeurs pour les sociétés, justement parce qu'elle véhiculerait la volonté de substituer des machines à l'homme et mettrait ainsi en péril certaines dimensions éthiques et matérielles de la vie humaine. Mobilisant aisément les fantasmes issus de la science-fiction, ces débats suscitent de manière plus générale des réflexions ontologiques sur les soubassements de l'intelligence humaine, de la technologie et des manières dont elles interagissent.

S'il n'existe aucune définition de l'IA qui fasse consensus, celle-ci peut néanmoins être assimilée à un ensemble de produits ou de services dont le fonctionnement dépend de programmes informatiques utilisant des technologies issues de l'IA. Ces produits peuvent être caractérisés en fonction de leur finalité ou de leurs fonctionnalités (prévision, traduction, reconnaissance…) ou du type de modèle qu'ils utilisent (apprentissage automatique, système expert…). Loin des fantasmes autour d'une « IA générale », c'est-à-dire capable de répliquer de façon autonome le comportement humain, ces produits concernent pour l'instant essentiellement une IA dite « étroite », c'est-à-dire capable de répliquer certaines tâches spécifiques, plus ou moins complexes, associées aux fonctions cognitives de l'humain (parler, écouter, percevoir…), grâce à une assistance humaine pour les concevoir ou les faire fonctionner. Il est néanmoins difficile de catégoriser de manière claire l'ensemble des objets concernés par l'IA, tant les entités

[1] Merci à Delphine Mantienne pour sa relecture attentive du texte.

techniques que cette dernière désigne tendent à évoluer rapidement (Cardon et Benbouzid, 2022). Mais l'IA est aussi, et sans doute en premier lieu, un champ de recherche qui mêle des sciences dures (les mathématiques et l'informatique), et des sciences sociales (linguistique, philosophie, sociologie, économie, droit et éthique) (Vayre et Gaglio, 2020). Ce champ de recherche alimente la recherche fondamentale permettant d'atteindre certains progrès scientifiques et nourrit la recherche appliquée en produisant des innovations technologiques mises sur le marché.

Les redéfinitions permanentes de ce qui constitue l'IA ont des conséquences importantes sur la manière dont le débat public s'en empare et surtout sur la manière dont chaque catégorie d'acteurs concernée souhaite la réguler (Cardon et Benbouzid, 2022). Comme le suggèrent Benbouzid *et al.* (2022) à propos des débats autour des enjeux de régulation de l'IA, « les problèmes définitionnels [sont] au cœur de conflits normatifs sur les moyens d'assujettir l'IA à un "contrôle social", qu'il soit technique, éthique, juridique ou politique ». Différents acteurs (entreprises du numérique, chercheurs en IA, associations, États…) proposent actuellement des mesures concrètes de régulation (chartes, lois…) qui interrogent la capacité des décideurs publics à la fois à intégrer des solutions IA pour élaborer, déployer et évaluer l'action publique, et plus généralement à réguler la diffusion de l'IA au sein des sociétés.

Ces réflexions quant à la place et aux définitions de l'IA ne sont pourtant pas nouvelles et semblent être des réactualisations sous des formes différentes de controverses plus anciennes. L'histoire de l'IA a pris son essor à partir de travaux scientifiques fondateurs dans les années 1940 et a été structurellement agitée par une opposition majeure entre deux courants scientifiques – l'IA symbolique et l'IA connexionniste – dont les promesses, les progrès mais aussi les espoirs déçus ont alimenté plus généralement les débats sur la place de l'IA dans la société. Cette histoire, sur laquelle nous revenons dans un premier temps, est riche d'enseignements car elle permet, d'une part, de bien délimiter les concepts fréquemment mobilisés pour parler d'IA (*deep learning*, IA générative, IA forte, IA faible…), et d'autre part, de saisir la spécificité des déploiements contemporains de l'IA. Compte tenu des développements majeurs que connaît l'IA actuellement, nous examinerons dans un deuxième temps les opportunités et les risques qu'elle présente pour l'action publique.

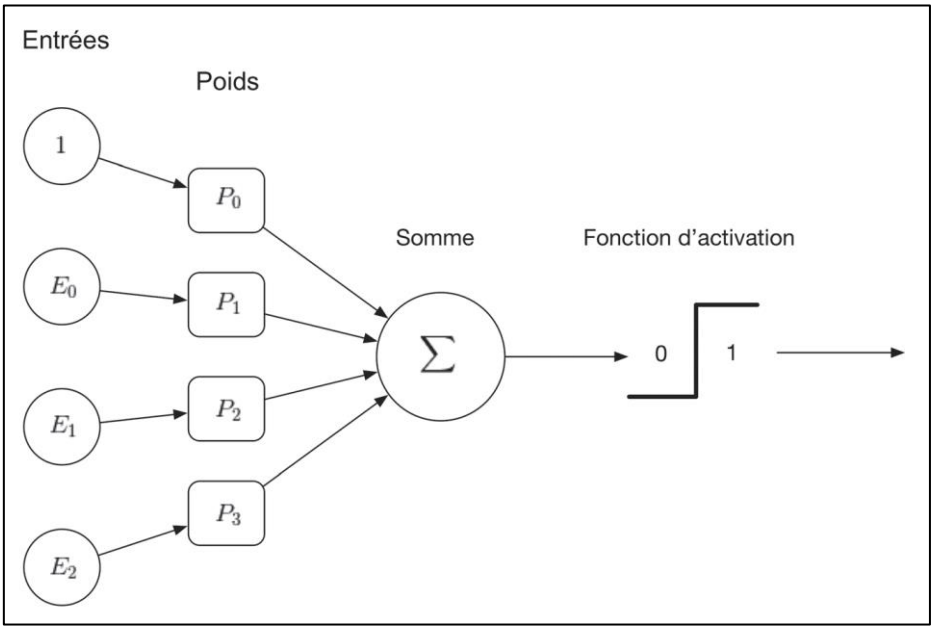
L'histoire en dents de scie de l'IA : des années 1940 jusqu'à nos jours

L'histoire de l'IA est avant tout une histoire des sciences et des techniques qui voit se succéder des périodes d'euphorie liées aux avancées scientifiques dans le domaine, et des périodes de redimensionnement qui marquent l'échec relatif des perspectives promises par les scientifiques. L'histoire de l'IA peut être considérée à travers le prisme des controverses entre deux courants scientifiques, les tenants de l'IA connexionniste et ceux de l'IA symbolique. Ces deux courants ont tout à tour connu succès et échecs. Ils recouvrent deux manières de conceptualiser l'intelligence artificielle et concomitamment de concevoir le fonctionnement du cerveau humain.

La cybernétique ou l'essor avant l'heure de l'intelligence artificielle connexionniste

L'histoire de l'IA débute dans les années 1940 avec l'article fondateur de l'informatique et de la première cybernétique signé du neurophysiologiste Warren McCulloch et du logicien Walter Pitts (1943). Cet article modélise mathématiquement un réseau de neurones où chacun « prend des variables en entrées, y applique un poids pour produire une somme qui, si elle dépasse un certain seuil, déclenche l'activation du neurone » (Cardon *et al.*, 2018) (voir schéma 1).

Schéma 1 : Un réseau de neurones simple



Source : Cardon *et al.* (2018)

Ce modèle, qui fonde l'IA connexionniste avant même que le terme d'IA ne lui soit appliqué, est rapidement associé par le neuropsychologue Donald O. Hebb (1949) à la notion d'apprentissage. Ses travaux montrent en effet que l'activation répétée d'un neurone par un autre, à travers une synapse donnée, augmente sa conductivité et peut être considérée comme un apprentissage. Les modèles de réseaux de neurones deviennent le cœur du calculateur des premières machines « intelligentes » (Dupuy, 2005). S'appuyant sur la théorie de l'information développée par Shannon (1948) qui envisage l'information comme un signal, plutôt que comme un code, les travaux fondateurs de la cybernétique considèrent ainsi que les machines sont de simples boîtes noires apprenantes où l'on incorpore des entrées à partir desquelles sont produites des sorties. Selon une logique inductive, l'apprentissage machine (*machine learning*) est défini par la machine elle-même, et non par l'humain, à partir des données qui y entrent et à partir desquelles elle infère des relations, des récurrences, des corrélations, des liens de proximité. C'est donc la machine qui déduit des règles de ces données. Les sorties sont ensuite comparées avec le monde réel, ce qui permet de mesurer un écart entre ce dernier et le comportement de la machine, et ainsi de rectifier le fonctionnement de la machine pour le rapprocher du réel. En 1948, Norbert Wiener, mathématicien américain et l'un des pères de la cybernétique, développe par exemple un dispositif visant à prédire le guidage des missiles anti-aériens. Cette machine met ainsi à jour en continu la trajectoire des missiles en comparant le trajet effectif vers la cible avec les précédentes estimations. Elle doit évoluer vers la meilleure solution en fonction des données disponibles qui nourrissent, corrigent et orientent le calculateur.

À la fin des années 1950, les réseaux de neurones vont connaître un développement important, notamment en matière de reconnaissance visuelle, suscitant une première vague d'intérêt public pour le domaine. L'initiative la plus emblématique de ces premières machines connexionnistes, qui se déploie entre 1957 et 1961, reste sans doute le *Perceptron* du psychologue et informaticien de l'université de Cornell Frank Rosenblatt, configuré pour la reconnaissance d'images. Financé par la marine américaine (*Office of Naval Research*), le *Perceptron* met en œuvre des neurones d'entrée qui simulent l'activité de la rétine, et des neurones de sortie qui classent les « traits » reconnus par le système. La structure du réseau est alors automatiquement organisée par le mécanisme d'apprentissage statistique de la machine. Bien que Franck Rosenblatt ait pu entamer la construction d'une première version du Perceptron, le Mark I, le déploiement de plus ample envergure se heurte rapidement aux limites techniques de l'époque et à la concurrence naissante d'autres courants scientifiques, en particulier l'IA symbolique.

L'IA symbolique, ou la naissance de la « vraie » intelligence artificielle

Le terme d'« intelligence artificielle » est officiellement inventé par les chercheurs John McCarthy et Marvin Minsky lors d'une célèbre réunion à Dartmouth en 1956. Ce terme est sciemment utilisé pour s'opposer à la première cybernétique (Dupuy, 2005) et proposer une véritable théorie de l'esprit qui permette de programmer des règles répliquant les décisions humaines au cœur des machines. L'approche symbolique de l'IA dont McCarthy et Minsky sont les précurseurs va devenir largement majoritaire dans les travaux scientifiques des années 1960 jusqu'au début des années 1990. Ces travaux vont être particulièrement médiatisés jusqu'au début des années 1970, notamment ceux d'un petit groupe de chercheurs du MIT (Minsky, Papert), de l'université de Carnegie Mellon (Simon, Newell) et de Standford (McCarthy) qui, entre 1964 ET 1974, vont recevoir 75 % du financement des recherches en IA distribué par l'ARPA[2] et l'Air Force (Fleck, 1982). Cette montée en puissance des travaux d'IA symbolique s'explique par des attaques répétées visant à décrédibiliser les travaux cybernéticiens et dont l'objectif attend son but à la mort de F. Rosenblatt en 1971, date à laquelle les recherches sur les réseaux de neurones cessent d'être financées.

L'IA symbolique se distingue ainsi de l'IA connexionniste car elle suppose que la programmation permet de doter les machines de capacités abstraites et logiques, autrement dit d'une capacité de raisonner (Gardner, 1985). Basée sur des règles, elle reproduit un raisonnement logique de type hypothético-déductif (« si x et y, alors z »). Les règles qui représentent le savoir dans un domaine d'application sont stockées dans une base de connaissances. Un moteur d'inférence sélectionne dans cette base les règles pertinentes et les applique pour résoudre le problème posé, en fonction des faits qui lui sont soumis. Naissant la même année que les sciences cognitives (1956), l'IA symbolique se présente aussi comme une théorie computationnelle de l'esprit qui suppose que les états mentaux « peuvent être décrits à la fois sous forme matérielle comme un ensemble de traitements physiques des informations et sous forme symbolique comme des opérations, exécutables mécaniquement, de comparaison, de hiérarchisation ou d'inférence sur des significations (Andler, 2016) » (Cardon *et al.*, 2018). La conception symbolique de l'IA est alimentée par les progrès techniques des ordinateurs qui, dans les années 1950, séparent les opérations logiques effectuées sur des symboles (*software*) de la structure matérielle de la machine (*hardware*), selon une nouvelle configuration dite de Von Neumann (1945). Ces avancées permettent de créer un espace propre où il est possible de programmer des règles avec un langage humain, ensuite transcrites en langage machine sous la forme de 0 et de 1, indépendamment du fonctionnement de l'ordinateur. Des applications à certains domaines, notamment les « problèmes jouets » comme les jeux d'échecs ou de dames (Samuel, 1952 ; 1959) ou des théorèmes de géométrie (Gelertner, 1959), serviront de terrains d'expérimentation à ces travaux. La machine est alors paramétrée pour respecter les règles du jeu et adopter des stratégies pré-programmées en fonction des mouvements du joueur humain.

En somme, « l'intelligence artificielle peut désormais se penser comme une science de l'esprit dans la machine » (Cardon *et al.*, 2018). L'IA développée est alors revendiquée comme « forte » car les objectifs donnés aux machines leur sont propres et peuvent être déduits d'une forme de raison incorporée dans les inférences opérées par les programmes. À partir de la programmation incorporée dans la machine, celle-ci doit résoudre le problème, trouver la solution vraie ou correcte et prendre la décision satisfaisante. Néanmoins, les promesses des promoteurs des deux formes d'IA – symbolique et connexionniste – ne rencontrent pas les résultats escomptés, en dépit d'avancées manifestes. Dans le contexte de la guerre froide, F. Rosenblatt avait par exemple promis à la marine américaine qu'elle serait prochainement dotée d'un ordinateur capable de marcher, parler, voir, écrire et être conscient de son existence. De leur côté, Herbert Simon et Marvin Minsky avaient promis des traducteurs de textes russes, des robots infiltrés dans les lignes adverses et des systèmes de commande vocale pour les tanks et avions. Au début des années 1970, les financeurs publics, essentiellement

[2] L'ARPA correspond à l'ancienne *Defense Advanced Research Projects Agency* (DARPA), l'agence du département de la Défense des États-Unis chargée des nouvelles technologies militaires.

militaires, coupent alors leurs financements, plongeant l'IA dans un premier hiver (Cardon *et al.*, 2018).

La deuxième vague de l'IA : les systèmes experts

Grâce aux progrès des calculateurs, rendus plus puissants, l'IA symbolique retrouve néanmoins des couleurs au début des années 1980 et ses tenants se lancent dans une profonde révision de l'architecture des machines. Cette efficacité accrue des ordinateurs permet aux chercheurs de faire entrer des informations bien plus nombreuses dans la mémoire des ordinateurs et de dépasser les limites des années 1970 où les machines symboliques échouaient à rendre des résultats satisfaisants dès que l'univers modélisé se complexifiait. Les chercheurs symboliques vont alors intégrer des répertoires de connaissances spécialisées issus des savoirs d'experts. Plus précisément, la nouvelle logique à l'œuvre consiste à interagir avec un monde extérieur, non conçu par les programmateurs, et à transformer les connaissances détenues par des spécialistes de différents domaines en langage informatique afin que les utilisateurs des systèmes experts puissent leur poser des questions. Cette évolution est permise par la transformation des architectures des machines qui séparent le « moteur d'inférence » et une série de *mondes* possibles appelés « systèmes de production » selon les termes d'Edward Feigenbaum, créateur de DENDRAL, premier système expert d'identification des composants chimiques des matériaux (Cardon *et al.*, 2018). Comme le précisent Cardon *et al.* (2018), « les données qui nourrissent ces bases de connaissances consistent en de longues listes facilement modifiables et révisables de règles du type "SI… ALORS" (par exemple : "SI FIÈVRE, ALORS [CHERCHER INFECTION]") qui sont dissociées du mécanisme permettant de décider quand et comment appliquer la règle (moteur d'inférence) ».

Progressivement, les règles de raisonnement utilisées par les machines sont par ailleurs probabilisées, ce qui permet de relativiser le raisonnement inférentiel des premières machines symboliques. Autrement dit, au lieu de se référer au couple « vrai/faux », les nouvelles machines symboliques y substituent des estimations du caractère correct et pertinent des réponses données par les systèmes. La pertinence de ces réponses est ensuite évaluée par les experts pour raffiner l'apprentissage artificiel des machines. Dans les années 1980, les systèmes experts vont être développés dans les milieux scientifiques et industriels, par exemple pour configurer des ordinateurs, pour identifier des pannes de locomotives ou pour repérer des gisements géologiques (Crevier, 1997). Malgré tout, les systèmes experts connaissent une complexification croissante qui nécessite de créer sans cesse des répertoires de règles explicites. Leur mise en œuvre semble d'autant plus illusoire qu'elle implique de s'équiper de stations de travail extrêmement chères, au moment même où le marché des PC commence à se développer. Au début des années 1990 débute un deuxième hiver pour les chercheurs en IA, alors même qu'en parallèle, les travaux connexionnistes, bien qu'isolés, n'ont eu de cesse de se réinventer.

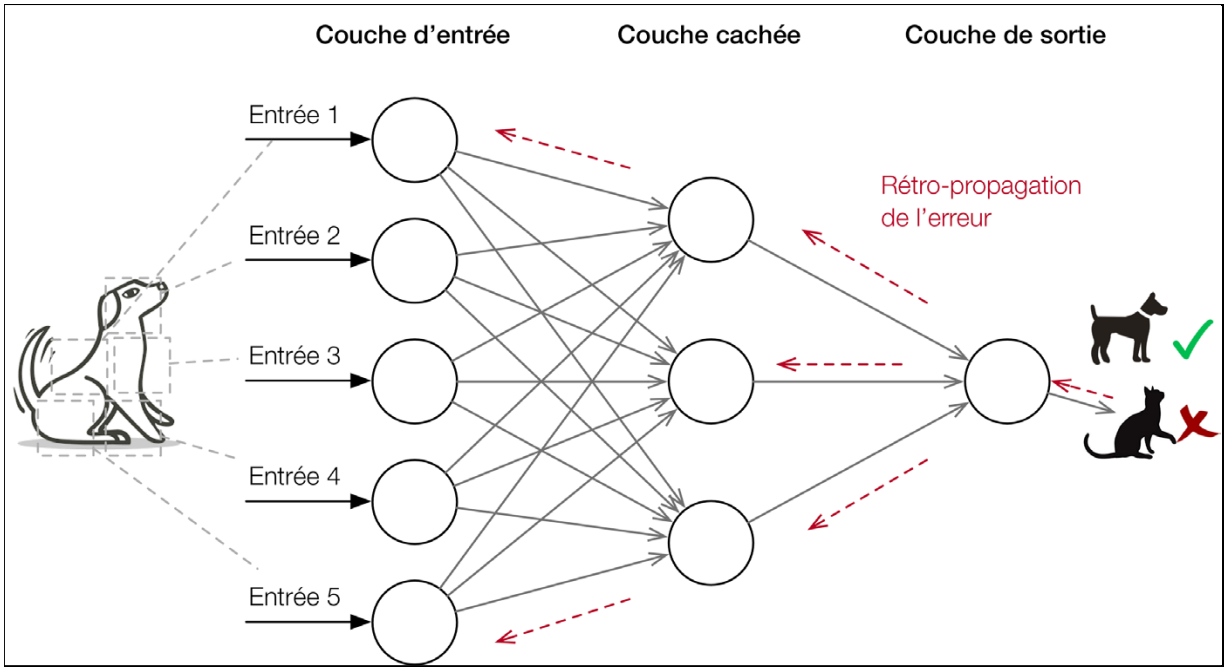
Le* deep learning *et le renouveau des approches connexionnistes

Bien que mis au ban, les travaux connexionnistes n'ont pas complètement disparu entre les années 1970 et 1990. Au contraire, à partir des années 1980, ils perdurent autour d'une importante créativité théorique et algorithmique qui cherche à identifier des artifices pour améliorer les réseaux de neurones et les coefficients qui lient entrées et sorties des machines. Ils prennent un tournant décisif à partir de la conception d'un algorithme – la rétropropagation de gradient stochastique (« *backprop* ») – qui permet de calculer le poids des coefficients (Rumelhart *et al.*, 1986a). Les auteurs de cet article dotent le réseau de neurones de plusieurs couches, ce qui permet de l'entraîner de façon simple[3] (voir schéma 2). Les intuitions de la première cybernétique sont alors rendues possibles par ces méthodes dans la mesure où ce

[3] Comme l'expliquent Cardon *et al.* (2018), « les couches additionnelles de neurones permettent d'apprendre des fonctions non linéaires. L'algorithme fonctionne en prenant la dérivée de la fonction de perte du réseau et en "propage" l'erreur pour corriger les coefficients dans les couches basses du réseau – dans un esprit proche des machines cybernétiques, l'erreur en sortie est "propagée" vers les entrées. »

fonctionnement à plusieurs couches permet de faire remonter l'erreur en sortie vers les coefficients qui sont alors autocorrigés pour résoudre l'erreur.

Schéma 2 : Le raffinement de l'IA connexionniste



Source : Cardon *et al.* (2018).

L'algorithme pouvant être généralisé à tout type de réseaux de neurones, la recherche connexionniste va s'engouffrer dans la brèche ouverte par ces travaux. Les réseaux de neurones artificiels, qui suscitent l'engouement le plus important actuellement, vont progressivement être dotés d'au moins trois couches de neurones : une couche d'entrée qui reçoit des données brutes ; reliée à une couche cachée qui traite ces données ; elle-même reliée à une couche de sortie qui produit le résultat. Lorsque le réseau comporte plus d'une seule couche cachée, on parle alors d'apprentissage profond (*deep learning*). Une de ses premières applications concerne par exemple la reconnaissance réussie des codes postaux de la poste américaine opérée par Lecun *et al.* (1989) grâce à un réseau multicouches capable, à partir des données de l'US Postal Service, de détecter les chiffres du code postal indiqués sur les colis.

Le développement parallèle des entreprises du web, qui n'hésitent pas à emprunter des techniques d'apprentissage artificiel dans leurs analyses des données (*datascience*) structurant nombre de leurs activités (détection des spams, recommandations pour les utilisateurs, analyse des réseaux sociaux…), va permettre de confirmer ce renouveau connexionniste. En collaboration avec les acteurs du numérique, notamment Google, Facebook et Baidu, le courant connexionniste va utiliser les *big data* que ces acteurs leur transmettent pour mettre à l'épreuve du réel l'efficacité de la prédiction de leurs machines. La multiplication des données d'entraînement des machines permet que les représentations apprises par le réseau deviennent robustes et tolérantes aux erreurs dans les données d'apprentissage. De 9 298 chiffres manuscrits de codes postaux dans l'article de Lecun *et al.* (1989), les *datasets* vont alors être radicalement augmentés, atteignant fréquemment plusieurs millions de données dans les travaux actuels.

Au-delà du nombre, c'est aussi la manière dont les données d'entrée vont être traitées qui évolue progressivement grâce à des opérations dites de « plongement » (*embedding*). Ces opérations permettent de coder les données sous la forme de représentations vectorielles purement numériques. Elles seront particulièrement développées pour des objets complexes, dont le texte qui en constitue l'exemple prototypique. Comme l'expliquent Cardon *et al.* (2018) à propos du codage textuel, « pour faire entrer un mot dans un réseau de neurones, la

technique Word2vec le "plonge" dans un espace vectoriel qui mesure sa distance avec tous les autres mots du corpus (Mikolov *et al.,* 2013). Les mots héritent ainsi d'une position dans un espace de plusieurs centaines de dimensions. […] Deux termes dont les positions inférées dans cet espace sont proches sont également similaires sémantiquement, on dit de ces représentations qu'elles sont distribuées : le vecteur du concept "appartement" [-0.2, 0.3, -4.2, 5.1…] sera proche de celui du "maison" [-0.2, 0.3, -4.0, 5.1…] ». Les opérations vectorielles permettant de remplacer les mots comme données en entrées des machines se généralisent à un ensemble de tâches de plus en plus complexes, comme la classification automatique de documents ou le résumé automatique. À partir de la fin des années 2000, les progrès sont tels qu'ils atteignent l'ensemble des domaines historiquement concernés par l'IA : le signal, la voix, l'image et la traduction automatique. Dans la continuité, les calculs des réseaux de neurones se perfectionnent pour identifier des caractéristiques et reconnaître des formes dans les données initiales. L'une des illustrations les plus emblématiques de ce perfectionnement concerne le principe de convolution qui permet d'identifier les bords, les coins, les contrastes et des points particuliers dans des images et de les associer à des ensembles de mots liés aux objectifs confiés à la machine.

En somme, la « révolution connexionniste » à l'œuvre aujourd'hui se déploie « dans un monde dans lequel les données doivent non seulement être massives, mais aussi les plus atomisées possible afin de leur ôter toute structure explicite. Si les données enferment bien des régularités, des relations compositionnelles, des styles globaux, etc., ceux-ci doivent être mis en évidence par le calculateur et non par le programmeur. Le premier trait du travail de production de l'induction consiste donc à faire entrer dans le système des données sous la forme la plus élémentaire possible : des pixels plutôt que des formes, des fréquences plutôt que des phonèmes, des lettres plutôt que des mots, des clics plutôt que des déclarations d'internautes, des comportements plutôt que des catégories… » (Cardon *et al.*, 2018).

L'important développement de l'IA connexionniste a abouti aujourd'hui à plusieurs types d'apprentissages, plus ou moins étroitement liés à l'intervention humaine sur les données qui alimentent les machines.

- L'apprentissage supervisé lorsque les données ont été étiquetées : l'humain leur a assigné une valeur ou une catégorie, afin de permettre à la machine de les différencier (par exemple, *telle photo représente une chaise*). Le paramétrage « manuel » de la machine doit ainsi permettre que la prévision qu'effectue la machine (inférence) à partir de données d'entrée (n'importe quelle photo) soit conforme ou la plus proche possible de la valeur attendue (l'outil devra indiquer qu'une chaise se trouve sur la photo si tel est le cas et le résultat inverse sinon).
- L'apprentissage non-supervisé où l'IA se cantonne à structurer le jeu de données en les partitionnant ou en les segmentant (*clustering*) par catégories homogènes, afin d'identifier des différences, des récurrences ou des anomalies. Le système est incapable de qualifier chacune de ces catégories, mais il en identifie l'existence et un certain nombre de traits distinctifs. Un système basé sur l'apprentissage non supervisé peut, par exemple, classer les plantes en distinguant un ensemble d'images A (*des orchidées*) et un ensemble d'images B (*des roses*) en raison des similitudes respectives au sein du groupe A et au sein du groupe B, et de la dissemblance entre les spécimens de A et B.
- L'apprentissage auto-supervisé qui se situe entre les deux types d'apprentissages précédents et qui s'avère prometteur pour contourner le problème de la rareté des données étiquetées ou du coût de leur étiquetage. Il consiste à produire automatiquement des données d'entraînement, notamment en altérant un jeu de données afin que la machine retrouve le modèle initial. Il peut s'agir, par exemple, de tronquer des phrases (en enlevant un mot) ; le modèle doit alors « deviner » le mot manquant, auquel cas il est validé ou corrigé sinon.
- L'apprentissage par renforcement, qui consiste à définir un objectif dont l'atteinte entraînera une « récompense » pour la machine. AlphaGo Zero, version avancée du programme AlphaGo initialement basé sur l'apprentissage supervisé et ayant battu pour la première fois le champion du monde du jeu de go, a par exemple été entraîné de cette manière. Aucune donnée issue de parties jouées par des champions humains

n'a alors été mobilisée et l'humain n'est intervenu que pour intégrer les règles du jeu et le système de récompense. Le système a alors joué contre lui-même pendant une durée déterminée jusqu'à ce qu'il atteigne des performances jugées satisfaisantes par l'homme.

Les progrès accomplis par l'IA, fruit de controverses multiples, souvent acerbes, entre les communautés symbolique ou connexionniste depuis plusieurs décennies semblent aujourd'hui aboutir à une victoire de l'IA connexionniste. Pourtant, la recherche actuelle s'orienterait vers une combinaison des deux approches plutôt que vers une exclusion d'un courant au profit de l'autre. Ces deux logiques sont d'ailleurs déjà combinées au sein de certains systèmes d'IA. Tel est le cas, par exemple, de certaines IA capables de détecter la présence d'un bâtiment sur une image satellite, le qualifier (maison, usine, piscine…), à l'aide d'un modèle d'apprentissage machine, puis, par l'application de simples règles, de déterminer certaines de ses caractéristiques (date de construction par exemple) ou d'apprécier certains éléments spécifiques (nécessité d'une délivrance d'une autorisation d'urbanisme, existence ou pas d'un permis de construire…).

Ce bref récit de l'histoire de l'IA rappelle d'une part que les promesses de cette dernière ont souvent abouti à des échecs retentissants et d'autre part que les architectures techniques qui la soutiennent, loin d'être neutres, doivent être rendues intelligibles pour que l'on comprenne son potentiel, tout autant que ses risques. En l'occurrence, en ce qui concerne les réseaux de neurones, si l'attention du régulateur doit être portée sur les stratégies des acteurs de l'économie du numérique qui les déploient, il convient également de veiller à la qualité des données qui entrent dans les machines comme à la manière dont les machines les utilisent pour faire leurs prédictions, et aux objectifs qui leur sont assignés.

Transformer l'action publique par l'IA ?

En répliquant certaines fonctions cognitives de l'humain (la perception, la mémoire, le raisonnement, la planification, l'apprentissage…), les systèmes IA pourraient profondément modifier le travail, l'organisation et l'action administratives. Au-delà des opportunités économiques, notamment en matière commerciale, industrielle et de compétitivité, qui ne concernent pas directement le management public, des leviers majeurs semblent se dessiner au moins dans deux directions : en matière d'amélioration du service public et en matière d'optimisation des ressources publiques (De Sousa *et al.*, 2019 ; Zuiderjwik *et al.*, 2021 ; Wirtz *et al.*, 2021 ; Bertolucci, 2024).

Vers un service public augmenté par l'IA ?

Différentes dimensions de la relation de service public pourraient être affectées positivement par l'intégration de solutions IA dans le fonctionnement administratif.

- Mutabilité et IA

La collecte et l'exploitation massives de données peuvent permettre d'acquérir, de se représenter et de gérer de plus amples connaissances sur les besoins des usagers, les activités économiques et sociales, l'environnement dans lequel l'action et les politiques publiques interviennent, ainsi que sur les effets de ces dernières. L'IA pourrait ainsi visibiliser et/ou représenter certains phénomènes méconnus ou inconnus, permettre de mieux expliquer leurs causes et conséquences et de calibrer des politiques publiques, tout autant que leur évaluation. Certaines fonctionnalités de l'IA déjà utilisées dans l'action publique en sont des illustrations concrètes. D'une part, la visualisation des données (*dataviz*), qui représente les données traitées sous une forme compréhensible (par des graphiques, des cartographies, des environnements virtuels…), est par exemple utilisée par les jumeaux numériques en matière de gestion immobilière publique (Fruquet, 2024) ou plus généralement dans la gestion des territoires et des réseaux (transport, mobilité, collecte et gestion des déchets, gestion de l'eau, prévention des risques naturels…). D'autre part, l'analyse dite « prédictive » tire de données

historiques les variables explicatives d'un phénomène (météorologique, sanitaire, délictuel…) afin d'identifier la probabilité de ses futures manifestations (Benbouzid et Cardon, 2018 ; Vayre, 2018). Ce type de dispositifs est notamment utilisé pour la recherche de fraudes, dans la santé, dans la régulation économique[4] et de manière plus controversée dans la police (Benbouzid, 2018).

Dans la même veine, la catégorisation et la représentation des données permises par l'IA peuvent améliorer l'accès à l'information des usagers, en mettant à leur disposition des informations générales sur le service public ou spécifiques à leurs demandes administratives, de manière rapide et accessible. Le *Government Digital Service*, l'agence du numérique de l'État britannique, utilise ainsi un réseau de neurones pour associer automatiquement des mots-clés au contenu des pages du site GOV.UK. Les fonctionnalités de l'IA peuvent aussi permettre d'alimenter en informations les évaluations de l'action publique, comme l'illustre l'outil d'apprentissage supervisé développé par l'ARCOM pour mesurer la présence de femmes dans les médias audiovisuels et veiller à leur juste représentation.

La qualité des prestations pourrait aussi être améliorée par le recours à des solutions IA. L'utilisation de la vision par ordinateur permet par exemple d'analyser des images fixes, des vidéos ou des objets afin d'identifier et de catégoriser des personnes et des choses à partir de leurs caractéristiques physiques (formes, proportions, couleurs, motifs…). Cette technique de reconnaissance biométrique et faciale des personnes est notamment utilisée dans les services publics de sécurité et de police pour repérer un suspect ou une victime à partir d'une photo ou d'une vidéo, ou en matière de sécurité routière pour les plaques d'immatriculations de conducteurs en infraction. Elle peut aussi permettre de déceler des anomalies, notamment en matière de fraudes, comme l'illustrent les initiatives de Tracfin, des douanes ou de l'administration fiscale (Cluzel-Métayer et Prebissy-Schnall, 2022). Depuis 2014, la DGFiP utilise par exemple des systèmes IA afin de programmer des contrôles fiscaux en fonction des données externes et d'informations passées, tout autant que pour repérer des bâtiments ou des aménagements non déclarés, notamment les piscines. L'IA peut être une aide pour les agents dans la conduite de leurs missions, autres que celles de contrôle, par exemple lorsque des outils IA sont utilisés par des professeurs d'école pour personnaliser les exercices en fonction du niveau et de la progression des élèves.

Le recours à l'IA pourrait également influencer positivement la vitesse de versement des prestations, en automatisant totalement ou partiellement certaines tâches administratives, qu'elles soient répétitives, comme l'envoi de courriers, ou plus spécifiques, comme la gestion des demandes d'allocations, en vérifiant et en classant automatiquement des pièces ou des informations transmises par les usagers. Des outils de traitement automatique des langues (TAL) ou traitement automatique du langage naturel (TALN) sont par exemple utilisés pour l'analyse sémantique de textes et l'identification, l'interprétation et le classement de leur contenu. La robotique immatérielle, qui simule le comportement humain dans un environnement numérique, peut quant à elle permettre de confier une tâche répétitive, comme la saisie ou la collecte de donnée, à une machine. Ces outils participeraient aussi d'une simplification administrative en développant une relation aux usagers par des interfaces simples d'utilisation et en aidant les agents publics, voire en se substituant à eux, pour mettre en œuvre les textes. L'utilisation d'outils IA, que ce soit dans le domaine social pour distribuer automatiquement des prestations à leurs bénéficiaires, éducatif pour affecter des élèves à des cycles de formation avec Parcoursup, ou judiciaire pour anonymiser des décisions juridictionnelles du Conseil d'État ou de la Cour de cassation, en sont autant d'illustrations.

- Continuité et IA

L'IA pourrait contribuer à rendre le service public plus continu. Le TALN peut par exemple permettre d'alimenter des robots conversationnels en générant des réponses-types à des

[4] L'outil « Signaux faibles », expérimenté en Bourgogne-Franche-Comté à partir de 2016 et étendu au niveau national en 2019, conçu par une collaboration entre la DGE, la DGEFP, l'Urssaf Caisse nationale, la Banque de France et la Dinum, cherche par exemple à détecter plus précocement les entreprises de plus de 10 salariés en difficulté à partir des données financières, économiques et d'activité détenues par les administrations.

questions plus ou moins simples, à n'importe quelle heure de la journée et de la nuit. Ces fonctionnalités sont particulièrement développées au sein des institutions directement en contact avec les usagers contribuables, allocataires ou cotisants. Le *chatbot* thématique « de crise » déployé par l'Urssaf Caisse nationale dans le cadre de la crise sanitaire a ainsi répondu à un million de questions simples, ce qui aurait été impossible en y affectant des agents humains aux horaires ouvrables. À l'étranger, l'outil Kommune-Kari, déjà disponible pour les habitants de 80 communes norvégiennes représentant plus de 30 % de la population, et dont la base de connaissances est l'une des plus larges au monde dans la sphère publique (6 000 entrées), est un robot conversationnel en cours de déploiement en Suède, en Finlande et au Danemark.

- Égalité et IA

L'IA pourrait enfin permettre une meilleure égalité de traitement entre usagers. D'une part, afin de rendre visibles ou d'améliorer l'étude de certaines inégalités invisibles jusqu'alors, grâce à la mise à nu de certains facteurs explicatifs présents dans les données. D'autre part, pour lutter contre le non recours en se substituant aux usagers et en étant proactive dans la recherche des bénéficiaires. Par exemple, depuis 2018, un algorithme développé par l'agence de la biomédecine établit l'ordre de proposition des greffons cardiaques, ressource vitale en quantité insuffisante pour les malades qui en ont besoin. Bien que l'attribution semble imparfaite, ce dispositif est jugé préférable à une attribution humaine qui était largement manipulée par les centres de greffe (Hénin, 2021).

Optimiser l'emploi des ressources publiques

Les systèmes IA peuvent être mobilisés pour améliorer l'utilisation des moyens matériels, pour des enjeux d'économies budgétaires ou pour poursuivre d'autres objectifs, liés par exemple à la transition écologique. La performance énergétique des bâtiments publics peut ainsi être modulée en fonction de l'exploitation des données historiques de consommation.

En outre, les systèmes IA peuvent être exploités en matière de GRH. Premièrement, un certain nombre de métiers administratifs vont être transformés, voire supprimés à mesure que l'automatisation sera perfectionnée (Conseil d'État, 2022 ; Premier ministre, 2024)[5]. Les types d'emplois concernés restent malgré tout encore difficiles à évaluer, ce qui pose la question d'effets différenciés en fonction des activités. Des études liminaires tendent à montrer que l'automatisation de tâches répétitives, là où l'IA est pour le moment la plus performante, touche en premier lieu les emplois à moindre valeur ajoutée, en améliorant leur productivité, laissant penser à une potentielle diminution des inégalités d'emploi (Brynjolfsson *et al.,* 2023. Noy et Zhang, 2023). Deuxièmement, en automatisant certaines tâches, l'IA pourrait permettre de dégager du temps administratif, notamment pour les dossiers les plus complexes nécessitant des moyens humains importants, par exemple dans la santé, le social, le médico-social, la police ou la gendarmerie et ainsi, de repositionner certains emplois sur la supervision et la maintenance des systèmes IA. Troisièmement, l'IA pourrait être mobilisée pour accompagner les services des ressources humaines en matière de recrutement[6], de gestion des parcours, de formation professionnelle continue (Lacroux et Martin-Lacroux, 2021). Elle permettrait entre autres d'automatiser l'information aux agents publics concernant leurs droits et leurs obligations, de leurs avantages sociaux ou de leurs formations. La Direction générale des douanes et des droits indirects (DGDDI) se penche depuis 2019 sur un simulateur de mutations chargé d'évaluer les probabilités d'être affecté à tel ou tel poste, en fonction d'une analyse automatisée du parcours des candidats.

[5] Par exemple, le rapport de la Commission de l'intelligence artificielle, commandé par le Premier ministre, « *suggère un effet positif de l'IA sur l'emploi dans les entreprises qui adoptent l'IA, car celle-ci remplace des tâches, et non des emplois. Dans 19 emplois sur 20, il existe des tâches que l'IA ne peut pas accomplir. Les emplois directement remplaçables par l'IA ne représenteraient donc que 5 % des emplois d'un pays comme la France.* » (p. 41).

[6] Le projet Tengai Unbiased développé en Suède prévoit par exemple d'associer un robot aux procédures de recrutement dans la fonction publique.

Encadrer les usages publics de l'IA

Relativement à ses potentialités, l'IA présente des risques liés à son déploiement au cœur de l'action publique. Outre des enjeux économiques et de sécurité (en matière de souveraineté industrielle et numérique, notamment), émergent des enjeux de gestion, qu'ils soient d'ordre technique, organisationnel ou éthique. Ces enjeux sont d'autant plus difficiles à appréhender du fait de ce que Collinridge (1980) appelle le dilemme du « contrôle social des technologies » caractérisé par un manque de connaissances pour prédire leurs conséquences dans la phase d'essor des technologies. Autrement dit, comme le précisent Cardon et Benbouzid (2022), « nous sommes contraints de laisser [ces technologies] se déployer pour en mesurer les conséquences, mais il est alors trop tard car elles sont enracinées dans la société et par un effet de dépendance leur contrôle devient difficile ».

Maîtriser les enjeux techniques de l'IA : un défi constant

Compte tenu de la complexité des systèmes utilisant l'intelligence artificielle, les risques techniques peuvent avoir plusieurs origines. Le calibrage des machines, de leur conception à leur utilisation, peut profondément affecter l'égalité ou la continuité du service public, comme la responsabilité des agents.

Des erreurs peuvent porter sur le paramétrage de l'algorithme qui réalise la tâche demandée. Les métriques et le seuil final choisis influencent fortement la qualité du système. Aux États-Unis, des algorithmes d'évaluation de la performance des enseignants ont été vivement critiqués car fondés uniquement sur les notes des élèves comme preuve directe de la performance. Par ailleurs, la qualité des données d'entraînement est cruciale pour la fiabilité des apprentissages. Enfin, l'éventuelle (contre)performance technique des machines est dépendante des autres dispositifs techniques auxquels l'IA est connectée pour fonctionner, par exemple lorsqu'un système de détection dépend de la bonne résolution des caméras de surveillance.

Les potentiels risques techniques rappellent qu'un usage pertinent des systèmes d'IA par l'action publique reste dépendant de l'action humaine, qu'il s'agisse de les configurer, de les contrôler, d'en penser les conditions d'utilisation ou de les réguler. Ils induisent nécessairement de réfléchir à la manière dont les systèmes IA peuvent être articulés avec les compétences requises pour les intégrer dans l'action publique, le travail administratif des agents et son organisation, tout en veillant à leur alignement avec les valeurs éthiques du service public.

Les compétences numériques comme enjeu critique du déploiement d'une IA publique

Le recours aux solutions IA dans l'action publique pose la question de l'adaptation des compétences présentes au sein de l'État aux opérations spécifiques de calibrage, de maintenance et d'utilisation des machines. En effet, les opérations de collecte, de sélection, de nettoyage et de qualification des données en vue de concevoir un système d'IA sont coordonnées par des ingénieurs spécialisés en IA, mais requièrent en outre un travail parfois déconsidéré d'annotation manuelle des données pour en identifier les éléments saillants à des fins d'entraînement des modèles. Ce travail du clic conditionne en grande partie les résultats produits et nécessite des ressources idoines, comme l'a montré Camille Girard-Chanudet (2023) à propos d'un outil d'anonymisation automatique des décisions de justice au sein de la Cour de cassation. Il peut notamment mobiliser des analystes de données (*data analysts*) qui participent aux tâches de formalisation et de représentation des données.

Le renforcement des compétences numériques de l'État est un autre enjeu pour les activités de contrôle. L'IA, notamment connexionniste, soulève des questions quant au fonctionnement des machines. En effet, les machines connexionnistes fonctionnent comme des boîtes noires et aucun spécialiste n'est encore capable d'expliquer le processus par lequel elles produisent leurs résultats. Cet état de fait implique qu'un contrôle humain, dans les phases de test, de

mise en production et de déploiement, est nécessaire pour s'assurer que l'IA respecte les objectifs qui lui sont assignés en matière de contrôle des administrés ou de relation à l'usager, et le fait dans le respect des valeurs publiques, notamment du point de vue de l'égalité de service public et des discriminations potentielles au sein des données.

La montée en compétence numérique est nécessaire pour que l'État soit capable de juger de la qualité des infrastructures techniques (processeurs, mémoire, stockage, services associés, super calculateur…) qu'elles soient développées en interne ou externalisées vers des prestataires privés. Un enjeu critique concerne le développement de ressources en matière de serveurs, de centres de données ou de *cloud* suffisamment performantes pour développer les activités publiques, tout en garantissant la confidentialité, la sécurité et la souveraineté des données. À travers la montée en compétences numériques publiques, ce sont donc aussi des défis de sûreté et de cybersécurité qui se font jour, pour anticiper et gérer d'éventuelles attaques informatiques, tout autant que la dépendance à des puissances extérieures pour faire fonctionner le service public.

En somme, la mise à niveau des compétences numériques agit aussi sur la capacité de l'État à proposer des services publics mobilisant l'IA crédibles et au moins aussi performants que leurs alternatives privées, et à ne pas se placer en situation d'asymétrie technologique, c'est-à-dire à pouvoir discuter d'égal à égal avec les prestataires privés.

Réorganiser le travail, former et acculturer les agents

Certains métiers, appelés à disparaître ou à être automatisés de façon massive, connaîtront des bouleversements. En termes de politique de l'emploi, même si les premières études montrent que les impacts de l'IA se distribueront de manière équitable entre emplois selon les niveaux d'étude et de revenu (Brynjolfsson *et al.,* 2023. Noy et Zhang, 2023), les observateurs envisagent d'ores et déjà que la politique publique aura à mettre en œuvre des dispositifs particuliers de gestion de la transition et de réinsertion professionnelle (Conseil d'État, 2022 ; Premier ministre, 2024).

Les agents publics qui demeureront en poste mais dont le travail sera accompagné par le recours à des solutions IA pour conduire leurs missions seront un autre enjeu. En l'occurrence, l'action publique devra être capable de mettre en œuvre une gestion prévisionnelle des compétences et des carrières adaptée à ce nouveau « choc » numérique afin, d'une part, de former et d'acculturer les agents et, d'autre part, de veiller à ce que cette intégration de l'IA dans l'exercice du travail administratif ne conduise pas à des effets pervers, comme un alourdissement de la charge du travail, une surcharge mentale liée aux irritants informatiques voire des résistances à l'usage des outils IA (Huang, 2022).

Il s'ensuit que l'intégration des outils IA dans la sphère publique doit prendre en compte les logiques métiers et professionnelles à l'œuvre. En analysant l'affaire ProPublica-Compas[7], Beaudouin et Maxwell (2023) montrent que les outils de prédiction de la récidive utilisés en justice pénale ne sont pas envisagés de la même manière par les *datascientists* qui en discutent le calibrage technique, et les acteurs judiciaires qui se les approprient en fonction de leurs pratiques professionnelles.

Intégrer l'éthique publique à l'IA ?

Très étroitement liés aux risques que soulèvent les usages de l'IA dans le secteur public, émergent des enjeux éthiques de différentes natures. Premièrement, l'autonomie humaine est tout à fait centrale. Certains métiers ou certaines tâches seront totalement ou partiellement

[7] Entre 2010 et 2016, aux États-Unis, le média ProPublica conduit une enquête sur l'outil COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), utilisé en justice pénale pour aider à quantifier les risques de fuite ou de récidive d'un prévenu, et rendre plus objectives les décisions de mise en détention provisoire ou de remise en liberté surveillée. À la fin de l'enquête, il révèle le caractère supposé raciste de l'outil, entraînant un débat public animé sur les usages des outils IA au sein de la sphère publique.

automatisées, questionnant le sens au travail que trouveront les agents publics dans l'agencement de leur activité avec celle des machines. C'est bien la question de l'autonomie des agents dans la conduite des politiques publiques qui est en jeu. Si de nombreux observateurs mettent en relief l'impact positif de l'IA en la matière, dans la mesure où elle permettrait d'automatiser un certain nombre de tâches parfois jugées comme redondantes et pénibles pour se consacrer à des tâches plus nobles, de réflexion notamment, d'autres montrent que cette assertion est peut-être plus complexe qu'elle n'y paraît (Bertolucci, 2023). Ils postulent notamment qu'une exposition constante aux IA pourrait au contraire diminuer nos capacités à créer, à apprendre et à penser, altérer les liens sociaux et l'autonomie des humains, en créant parfois des formes de dépendance ou d'affect vis-à-vis des machines. Par ailleurs, ce partage des tâches entre robots et humains interroge le partage des responsabilités entre les uns et les autres (Grozdanovski, 2022). Cette interrogation a des implications très pratiques dans la mesure où face à des décisions totalement ou partiellement automatisées, des usagers et/ou des agents publics les jugeant contraires à leurs droits pourraient se retourner contre elles, nécessitant pour l'administration de mettre en œuvre un régime de responsabilités clair.

Deuxièmement, le déploiement de l'IA s'accompagne d'un intérêt particulier pour la qualité des données ; qualité qui, dans le cas du secteur public, doit aussi rimer avec certaines valeurs publiques (non-discrimination, égalité…). En l'occurrence, un certain nombre de biais – de représentativité, sexistes, racistes, inégalitaires… –, parfois difficilement identifiables, peuvent se nicher dans les données, affectant les résultats de la machine (Beaudouin et Maxwell, 2024). Un rapport intitulé *Xenophobic Machines* (« Les machines xénophobes ») d'Amnesty international en 2019 avait par exemple montré que des critères relevant du profilage racial avaient été intégrés lors de l'élaboration de l'algorithme utilisé par l'administration néerlandaise pour évaluer si des demandes d'allocations familiales étaient erronées et potentiellement frauduleuses. Des dizaines de milliers de parents et de personnes ayant la charge d'enfants, souvent issus de familles modestes, ont été accusés à tort de fraude fiscale, les minorités ethniques étant touchées de manière disproportionnée. Il en résulte que l'action publique devra nécessairement être attentive à limiter ces biais au maximum, ce qui, compte tenu de la massification des données nécessaires au fonctionnement de ces machines, ne peut manquer de questionner. Et cela, d'autant plus que les possibilités de traduire des principes moraux en objectifs statistiques des machines semble être problématique, comme l'ont montré Beaudouin et Maxwell (2024) à propos de la mise en œuvre de critères d'équité dans des machines de prédiction de la récidive aux États-Unis.

Un point d'attention tout particulier concerne la plus grande personnalisation que l'IA pourrait rendre possible. En effet, l'IA, en assurant une meilleure segmentation et un meilleur profilage à partir des données, pourrait permettre aux services de mieux individualiser les relations avec les usagers, ce qui interroge la possibilité de faire coïncider cette personnalisation avec les logiques collectives structurant l'action publique (pluralisme démocratique et culturel, mutualisation du risque, égalité de traitement…) (Gueydier, 2020).

Le rôle central des données dans le fonctionnement des IA peut aussi se heurter à la protection des données personnelles (Cluzel-Métayer, 2022), et plus généralement aux libertés individuelles. Deux logiques sont être à l'œuvre : l'une, plutôt anglo-saxonne, considère que les données, à partir du moment où elles ont été transmises à une entité publique ou privée, appartiennent à cette dernière, et l'autre, consacrée dans la législation européenne (traduite en France dans le Règlement général de protection des données [RGPD]), considère que les données personnelles doivent être protégées par les organisations à qui elles ont été transmises.

Troisièmement, en lien avec l'observation précédente, la transparence apparaît comme un enjeu crucial pour le déploiement des IA dans le secteur public. D'une part, cette transparence concerne les données utilisées pour entraîner les IA dans leurs phases de test puis de déploiement. D'autre part, et plus généralement, c'est l'explicabilité des algorithmes et des machines qui est en jeu, c'est-à-dire la manière dont les machines prennent leurs décisions. Il

s'agit en ce sens de ménager des possibilités pour les autorités de contrôle, mais aussi pour les citoyens dans leur ensemble, à la fois d'être informés qu'une IA est mobilisée dans la relation qu'ils ont avec l'administration et de disposer des informations sur le fonctionnement du système. À ce titre, le gouvernement britannique a par exemple lancé à l'automne 2021 un des premiers standards nationaux de transparence algorithmique pour les organisations publiques. Construit à partir de délibérations publiques et à l'issue de réunions de fonctionnaires de plusieurs pays, de chercheurs et de spécialistes du secteur, son but est d'apporter des réponses concrètes au public sur le processus décisionnel et les modalités de l'utilisation des algorithmes dans l'action publique. Certaines études ont toutefois montré que cet enjeu d'explicabilité de l'IA était parfois relégué derrière d'autres priorités organisationnelles, notamment en lien avec la performance ou la capacité de rendre des comptes (Vuarin et Steyer, 2024).

En dernier lieu, la généralisation de l'IA dans le fonctionnement de l'économie, et en l'occurrence au sein de l'administration, contient le risque d'aggraver la crise environnementale, à la fois du fait de l'extraction de terres rares nécessaires aux machines et de la consommation d'électricité induite, par exemple, pour stocker les données (Le Goff, 2023 ; Sénat, 2020).

En somme, l'ensemble de ces enjeux, notamment éthiques, met en exergue la question de ce que certains spécialistes du champ de l'IA nomment l'alignement ou le problème de l'alignement, déjà discuté par l'un des pères de la cybernétique des années 1950 Norbet Wiener, à savoir la capacité des concepteurs et des développeurs à produire des IA dont les résultats s'orientent vers les objectifs que les humains souhaitent leur donner. Comme le précisent Cardon et Benbouzid (2022), « l'éthique de Wiener repose sur un triptyque qui n'a jamais cessé de structurer les débats autour de ce que nous pourrions appeler le Triangle d'or du contrôle de l'IA : protéger la frontière menacée entre les êtres vivants et les machines, ce qui suppose une prise de position ontologique sur la spécificité humaine vis-à-vis des intelligences artefactuelles ; toujours viser la fabrication de prothèses qui émancipent les hommes en lieu et place de la figure de l'"usine mécanisée" qui les remplace, les dépasse et plus insidieusement les "désengage" (Pickering, 2019) ; responsabiliser les scientifiques non pas de la maîtrise de leur "savoir-faire", mais du contrôle de leur "savoir quoi", c'est-à-dire de ce que doivent être nos buts à fabriquer des intelligences artificielles (Wiener, 1950). Inscrits dans différentes arènes, les multiples débats et critiques (Garvey, 2021) interrogeant les limites qui doivent être fixées au développement de l'IA peuvent être regardés comme des approfondissements de chacun des trois aspects de ce Triangle d'or ».

Si cet alignement se joue dans les détails techniques des machines, c'est aussi par un débat démocratique, impliquant potentiellement des usagers, des associations et plus généralement la société civile, que les objectifs et les finalités à donner aux IA pourraient être discutés. D'où la question de la bonne échelle de régulation, entre principes généraux et collectifs et adaptations plus locales. En radiologie, le champ pionnier où la question de la régulation des outils IA s'est posée, Mignot et Schultz (2022) ont montré qu'en dépit de principes abstraits présents dans des chartes et des rapports, la régulation s'est opérée au niveau des acteurs du domaine, à savoir les radiologues et les industriels du secteur. Il en résulte que la régulation qui a émergé s'est d'abord construite « autour des délimitations du groupe professionnel des radiologues et de la compétition entre les constructeurs historiques de dispositifs d'imagerie et les nouveaux entrants de l'innovation numérique » (Mignot et Schultz, 2022). L'un des enjeux essentiels de l'avenir de l'IA au sein du secteur public sera ainsi l'adaptation des principes généraux érigés dans les nombreux rapports récents sur l'IA et dans la législation à la variété des éthiques et des identités professionnelles qui composent l'administration, comme l'ont montré Gaglio et Loute (2023) dans une autre étude sur l'introduction d'outils IA en sénologie et en traumatologie.