# IA & normes

Nathalie Nevejans, Grégory Bonnet, Gaël Lejeune, Dominique Longin

AfIA

Association française
pour l'Intelligence Artificielle

# Bulletin N° 120

*Association française pour l'Intelligence Artificielle*

AfIA

AfIA
Association française
pour l'Intelligence Artificielle

# PRÉSENTATION DU BULLETIN

Le Bulletin de l'AfIA vise à fournir un cadre de discussions et d'échanges au sein des communautés académique et industrielle. Ainsi, toutes les contributions, pour peu qu'elles aient un intérêt général pour l'ensemble des lecteurs, sont les bienvenues. En particulier, les annonces, les comptes rendus de conférences, les notes de lecture et les articles de débat sont très recherchés.
Le Bulletin contient également chaque trimestre un dossier plus substantiel qui porte : soit sur un thème lié à l'IA (2 numéros par an), soit sur des équipes de recherche en IA (1 fois par an), soit sur la Plate-forme Intelligence Artificielle PfIA (1 fois par an).
Le comité de rédaction se réserve le droit de ne pas publier des contributions qu'il jugerait contraire à l'esprit du bulletin ou à sa politique éditoriale. En outre, les articles signés, de même que les contributions aux débats, reflètent le point de vue de leurs auteurs et n'engagent qu'eux-mêmes.

## Édito

Ce numéro du Bulletin de AfIA est consacré à un dossier pluridisciplinaire monté par Nathalie NEVEJANS, maîtresse de conférence en droit (Chaire IA Responsable, Université d'Artois) sur le thème « IA & Normes ». La question de l'ouverture de l'AfIA à la pluridisciplinarité avait été posée lors de l'AGO 2021 et ce bulletin est l'occasion d'aller dans ce sens.

Ce dossier met en lumière des questions récentes et cruciales pour l'Intelligence Artificielle. En effet, la question de la réglementation et de la régulation des systèmes d'intelligence artificielle devient de plus en plus récurrente, et un éclairage venant de collègues d'autres disciplines – en particulier le droit et les sciences humaines – nous semble particulièrement intéressant pour notre communauté. L'AfIA tient à remercier toutes les personnes qui ont alors contribué à la richesse de ce dossier (voir page 5).

Dans la suite de ce Bulletin, vous retrouverez le compte rendu de la Nuit de l'Info à laquelle le Collège CECILIA de l'AfIA a participé (voir page 102). L'AfIA félicite chaleureusement les équipes gagnantes de ce défis, à savoir « Codelab » et « Les Dodos Insomniaques, le retour ». Vous trouverez ensuite la liste des thèses soutenues lors du trimestre écoulé, page 106. Enfin, la composition actuelle du Conseil d'Administration se trouve en quatrième de couverture de tous nos bulletins.

Encore un grand merci à tous les contributeurs et contributrices de ce numéro, sans oublier Gaël LEJEUNE pour sa relecture assidue.

***Bonne lecture à tous !***

*Grégory BONNET*
*Rédacteur*

AfIA
Association française
pour l'Intelligence Artificielle

# SOMMAIRE

## DU BULLETIN DE L'AfIA

AfIA
Association française
pour l'Intelligence Artificielle

# Dossier

# « IA & Normes »

Dossier réalisé par

**Nathalie NEVEJANS**
*Maître de conférences HDR en droit*
*Titulaire de la Chaire IA Responsable (Droit, Éthique et Sciences de l'IA)*
*Université d'Artois*
*Directrice du DU Responsable de l'éthique de l'IA*
*nathalie.nevejans@univ-artois.fr*

# Édito

Un dossier spécial consacré au thème « IA & Normes » peut sembler intriguant, voire perturbant, de premier abord pour les chercheurs en IA. S'ils visualisent d'emblée de quoi il s'agit lorsqu'il est question d'IA, la notion de norme est moins évidente à appréhender, à juste titre.

Issu du latin *norma* qui signifie l'équerre, la règle ou la loi, le terme « norme » est polysémique. D'une manière générale, la norme se définit comme un état habituel, régulier ou conforme à la majorité des cas. Si on écarte tous les sens inutiles pour ce bulletin, on peut encore comprendre la norme de deux autres manières.

Les normes peuvent, d'une part, s'appréhender comme une prescription, sous la forme de règles ou de principes, qui se fonde sur des jugements de valeur afin d'agir ou de juger. Cette acception recouvre alors schématiquement plusieurs choses, toutes importantes pour ce dossier. On y trouve les normes juridiques, c'est-à-dire les règles de droit, émanant d'une autorité publique, qui régissent la vie des humains en société et qui s'imposent à tous. Les règles de droit ont pour spécificité d'être contraignantes et de fournir des outils coercitifs pour leur application effective en cas de non-respect. Les normes peuvent également se comprendre comme des normes morales ou sociales qui, bien que rarement sanctionnables au sens juridique du terme, sont loin d'être négligeables. Les normes morales se réfèrent à un ensemble de valeurs et de principes qui permettent de différencier le bien du mal, le juste de l'injuste, l'acceptable de l'inacceptable, et qui se traduisent par des prescriptions et des interdits auxquels se conformer. On passe de la morale à l'éthique dès lors que ces valeurs et principes moraux sont interrogés dans le cadre d'une démarche de réflexion argumentée en vue d'agir conformément à ceux-ci. Les normes sociales quant à elles sont davantage tournées vers le groupe en général et cherchent à assurer la cohésion du lien social en puisant dans les traditions. Même s'il s'agit de disciplines différentes avec des objectifs et des méthodes propres, les frontières sont poreuses. Ces différentes normes peuvent parfois se rapprocher, comme lorsque des normes éthiques sont à la source de normes juridiques, mais aussi s'éloigner totalement, comme lorsque des normes morales tentent de faire émerger des valeurs non admises par les normes sociales. Cependant, le point commun de ces normes juridiques, morales ou sociales est leur lien fort avec la culture qui les a vues émerger, de sorte qu'elles varient selon les régions du monde.

D'autre part, dans les domaines scientifiques et techniques, les normes correspondent aux exigences à respecter pour réaliser quelque chose ou pour mener à bien une étude. En matière industrielle ou technologique, cela recouvre une thématique également importante pour ce dossier. Il s'agit de la normalisation ou la standardisation, dont l'objectif est d'établir des référentiels de spécifications techniques pour certains produits, services et processus élaborés par un organisme national (Afnor en France), européen (CEN et ETSI) ou international (ISO) de standardisation. La grande différence avec les règles de droit est que ces référentiels de normes sont élaborés par des organismes de normalisation privés, et qu'ils sont le plus souvent d'application volontaire.

L'IA repose sur des technologies en évolution constante et rapide. Elle procure de nombreux avantages sur le plan économique, social ou sociétal. Comme la technique n'est pas neutre, ces bénéfices doivent être mis en balance avec les risques en termes humains, sociaux, sociétaux et environnementaux. C'est à cette immense tâche que les chercheurs des

sciences humaines et sociales (juristes, philosophes, sociologues, psychologues, etc.) se sont attelés depuis quelques années. En mesurant les enjeux et les défis présentés par l'IA, ils réfléchissent aux premiers jalons, aux premières normes, à faire émerger pour la société du futur. Les chercheurs en IA, et tous ceux qui sont impliqués dans la conception, le développement, la fabrication, la fourniture ou l'utilisation professionnelle d'un système d'IA, sont également concernés par les réflexions émergentes. En effet, partout dans le monde, la question de la régulation de l'IA se pose depuis quelques années. Si, préoccupées par les impacts de l'IA, certaines voies et voix – en général dans les contrées ultra-libérales – s'orientent vers une régulation non contraignante prenant alors seulement la forme de normes éthiques, c'est bien une régulation contraignante des usages de l'IA qui est en cours d'adoption dans l'Union européenne (UE). Afin d'encourager la confiance des citoyens européens dans l'IA, l'UE cherche à mettre en place une protection effective de la santé, de la sécurité et des droits fondamentaux des utilisateurs des systèmes d'IA, mais dans une approche équilibrée pour préserver l'innovation.

En réponse à l'appel lancé pour ce Bulletin de l'AfIA, seize contributions ont été reçues en provenance de la France et de nos voisins francophones, témoignant toutes de la conscience de ces préoccupations en matière d'IA. La norme aux multiples facettes entre en résonance avec le domaine de l'IA pour produire des réflexions multidisciplinaires, voire interdisciplinaires, riches, foisonnantes, variées et complémentaires. Ce numéro 120 regroupe ainsi les fruits des travaux de chercheurs et d'équipes de recherche, ou encore de divers experts, praticiens et spécialistes des SHS ou de l'IA qui travaillent sur, autour ou en lien avec la norme.

Je suis reconnaissante à l'Afia, et spécialement à son rédacteur-adjoint, Grégory BONNET, de m'avoir donné carte blanche pour élaborer ce dossier « IA & Normes ». Je remercie également les contributrices et contributeurs pour leur confiance. J'ai pris beaucoup de plaisir à vous lire et je sais qu'il en sera de même pour les lectrices et lecteurs de ce numéro.

## Les enjeux européens des normes techniques pour les systèmes d'IA

Par **Marion HO-DAC**
*CDEP UR 2471*
*Université d'Artois*
*marion.hodac@univ-artois.fr*
*http://cdep.univ-artois.fr/*

### Contexte

L'Union européenne (ci-après « UE ») prépare l'adoption prochaine d'un cadre réglementaire horizontal applicable aux systèmes d'IA dans le marché européen. La proposition européenne de loi sur l'intelligence artificielle (ci-après « *AI Act* ») est actuellement en négociation au sein des instances législatives de l'UE [2].

L'*AI Act* prend appui sur une classification des systèmes d'IA en fonction de leurs risques pour la santé, la sécurité et les droits fondamentaux (*i.e.* inacceptables, élevés ou faibles) et il ajuste, en fonction, les dispositions légales à appliquer par les fournisseurs desdits systèmes. Certains systèmes d'IA seront prohi-

bés, en particulier dans le domaine de l'identification biométrique et de l'évaluation des personnes physiques (*social scoring*) [1]. Quant aux systèmes à haut risque, notamment utilisés dans le domaine des infrastructures critiques, de l'éducation ou de la justice, ils seront soumis à des exigences essentielles (e.g. gestion des risques, gouvernance des données, transparence, contrôle humain, etc.) [2] qui devraient être traduites dans des « normes harmonisées » [3].

En ce sens, l'*AI Act* confère un rôle clé à la normalisation : il s'agit de l'adoption de prescriptions techniques non contraignantes auxquelles des produits, des procédés de fabrication ou des services peuvent se conformer et qui sont élaborées par consensus au sein d'organismes de normalisation [1]. Ainsi, les futures normes, d'adoption purement volontaire pour les organisations, permettront de mettre en œuvre concrètement les exigences essentielles applicables aux systèmes d'IA à haut risque, dès leur conception et tout au long de leur cycle de vie.

Le travail d'élaboration de ces normes européennes applicables aux systèmes d'IA à haut risque est en cours au sein des organismes européens de normalisation – le Comité européen de normalisation (CEN) et le Comité européen de normalisation électrotechnique (CENELEC) –. Il est important, selon nous, que les praticiens et chercheurs en IA, toutes sous-disciplines confondues, en soient informés et y participent.

## La portée juridique des normes harmonisées selon l'AI Act

L'*AI Act* prend appui sur le règlement (UE) n° 1025/2012 relatif à la normalisation [1] pour définir, dans son article 40, les contours juridiques qui seront donnés aux normes harmonisées. Cet article prévoit que les systèmes d'IA à haut risque conformes à des normes harmonisées (en totalité ou en partie) adoptées aux fins de traduire techniquement les exigences essentielles précitées de l'*AI Act* et dont la référence a été publiée au Journal Officiel de l'Union européenne, seront présumés respecter la réglementation contraignante de l'UE. Cette présomption de conformité confère donc une dimension « quasi-réglementaire » aux normes, ce qui témoigne de l'enjeu important de leur élaboration [4]. Cela suscite d'ailleurs des critiques sous l'angle de leur (manque de) légitimité constitutionnelle [5]. Les normes ne sont pas adoptées selon un processus démocratique ouvert et ne sont pas librement accessibles. Soumises à des droits de propriété intellectuelle, leur accès est en principe payant.

Les futures normes devront préciser et expliciter leur couverture exacte desdites exigences (par ex. en matière de gestion des risques, de qualité des données, d'enregistrement automatique des évènements, de transparence, de contrôle humain ou encore de robustesse, etc.), afin que le champ d'application de la présomption de conformité soit clairement établi. La norme harmonisée ne remplace donc pas le droit « dur » (en l'occurrence les dispositions de l'*AI Act*) mais constitue, pour l'opérateur, un moyen technique permettant de

1. Article 5 de la proposition *AI Act* [2].
2. Chapitre 2, articles 9 à 15 de la proposition *AI Act* [2].
3. Article 40 de la proposition *AI Act* [2].
4. Sous le contrôle de la Cour de justice de l'UE. Cette dernière a eu l'occasion de se prononcer sur la « justiciabilité » des normes harmonisées ; selon elle, ces dernières font partie du droit de l'Union [3]. En outre, dans le cadre de la procédure d'objections formelles à une norme harmonisée prévue par le règlement 1025/2012 sur la normalisation [1], le Tribunal et la Cour de justice peuvent connaître de recours en annulation contre de telle norme.

s'y conformer. Elle devrait permettre de réduire les risques du système d'IA concerné, sans décharger le fournisseur de sa responsabilité. En pratique, néanmoins, la réalité de cette responsabilité et les conséquences juridiques qui pourraient en découler dépendront de l'effectivité du cadre d'exécution de la future réglementation européenne des systèmes d'IA en matière de contrôle par les autorités publiques nationales, dont les autorités de surveillance du marché et le cadre de gouvernance régulatoire (*public enforcement*).

## Le cadre juridique européen d'élaboration des normes harmonisées en IA

L'élaboration des normes harmonisées au niveau européen prend juridiquement appui, de manière générique, sur une demande de normalisation de la Commission européenne adressée aux organismes européens de normalisation, sur le fondement de l'article 10 du règlement relatif à la normalisation [1]. Un projet de demande de normalisation dans le domaine des systèmes d'IA a été publié en décembre dernier par la Commission [4], ci-après « projet de demande de normalisation », et la version définitive devrait l'être en avril 2023.

Ce document fixe une série d'objectifs que les futures normes devraient viser, en particulier la défense des intérêts publics européens (*i.e.* santé, sécurité et respect des droits fondamentaux) et des valeurs de l'Union européenne, tout en assurant la croissance et l'innovation dans le secteur de l'IA [4]. Cet équilibre particulièrement délicat ressort, au sein du projet de demande de normalisation, de la conjonction entre, d'un côté, la liste générique des normes attendues (*deliverables*) mentionnées en Annexe 1 et, de l'autre, des exigences essentielles (*requirements*) qui y sont associées, pour chaque livrable normatif, en appui de l'*AI Act*, en Annexe 2.

Pour l'heure, dix domaines sont cités comme devant donner lieu à des livrables normatifs, accompagnés de caractéristiques clés, en application du futur règlement : (1) gestion des risques, (2) gouvernance et qualité des jeux de données, (3) enregistrement et journaux automatiques (logs), (4) transparence et informations aux utilisateurs, (5) contrôle humain, (6) exactitude, (7) robustesse, (8) cybersécurité, (9) système de gestion de la qualité et (10) évaluation de la conformité [4].

## La mise en pratique de la normalisation de l'IA

CEN et CENELEC ont créé, au second semestre 2021, un comité technique joint (*Joint Technical Committee*), actuel JTC 21, pour conduire les activités de normalisation en IA. Il est le point de contact pour la Commission européenne ainsi que pour les autres organismes de normalisation actifs en Europe dans le domaine de la normalisation de l'IA. Au niveau français, c'est la commission nationale IA (CNIA) de l'Afnor, l'organisme français de normalisation, qui est le point de liaison et assure la participation des parties prenantes françaises aux travaux du JTC 21.

Le JTC 21 devrait en principe fournir les livrables normatifs en matière d'IA au plus tard au 31 janvier 2025[5]. Pour cela, il devra formellement accepter la demande de normalisation qui lui sera notifiée prochainement par la Commission et lui présenter un « programme de travail » indiquant les normes, les organismes techniques responsables et un calendrier pour l'exécution des activités de normalisation demandées[6]. Il est également nécessaire que le JTC 21 puisse justifier des efforts entrepris pour assurer, d'une part, la participation multipartite des parties prenantes de l'écosystème

5. Annexe 1 du projet de demande de normalisation [4].
6. Article 2 du projet de demande de normalisation [4].

de l'IA, spécialement les PME et les organisations de la société civile de l'UE, et, d'autre part, une expertise particulière dans le domaine des droits fondamentaux [7].

Dans ce contexte, les milieux universitaires et de recherche scientifiques en IA doivent être informés, voire sollicités, pour s'engager dans le suivi, le dialogue et la co-construction des futures normes européennes en IA. En pratique, plusieurs sous-groupes d'experts sont déjà au travail sur les aspects opérationnels des systèmes d'AI (notamment risque et conformité), l'ingénierie (par ex. traitement du langage naturel, gouvernance et qualité de données) et leurs aspects sociétaux (par ex. caractères de l'IA de confiance, *nudges* renforcés par l'IA, durabilité) [8]. Pour l'heure, ces groupes de travail sont majoritairement composés des représentants des grandes entreprises privées de la « tech », largement d'ailleurs d'origine non-européenne. Il est donc important que les opérateurs européens, y compris les PME et la société civile, à travers ses chercheurs notamment, s'expriment.

Cela serait particulièrement pertinent s'agissant d'une fabrique « européenne » de la régulation de l'IA ancrée dans les valeurs de l'UE au sens de l'article 2 du traité sur l'UE. Si des normes techniques en IA sont également élaborées au niveau international, au sein de l'ISO en particulier [9], elles ne sont pas toutes transposables *per se* au contexte européen. En effet, les futures normes européennes sont projetées par l'UE comme devant venir en soutien du respect, par les systèmes d'IA, des droits fondamentaux, garantis par la Charte de l'UE, et des valeurs de dignité humaine, liberté, démocratie et État de droit. Il s'agit d'un vrai défi technique, politique et juridique à relever.

À cette fin, l'écosystème de la normalisation européenne des systèmes d'IA doit être porté à la connaissance de l'ensemble de la communauté de l'IA.

## Références

[1] Règlement (UE) n° 1025/2012 relatif à la normalisation européenne, 25 octobre 2012.

[2] Proposition de règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (dit « AI Act »), COM/2021/206 final, 21 avril 2021. Voir la version actuelle de l'*AI Act* : Orientation générale du Conseil de l'Union européenne modifiant la proposition d'*AI Act*, document ST 15698 2022 INIT, 6 décembre 2022.

[3] CJUE, 27 octobre 2016, C-613/14, *James Eliott*, spéc. point 43.

[4] European Commission. Draft standardisation request to the European Standardisation Organisations in support of safe and trustworthy artificial intelligence, 5 décembre 2022.

[5] M. Eliantonio and C. Cauffman (Eds.). *The legitimacy of standardisation as a regulatory technique : A cross-disciplinary and multi-level analysis*. Edward Elgar Publishing, 9781789902952, 2020.

7. *Ibid*.
8. Voir la diapositive 4 de la présentation du 19/12/22 de S. Hallensleben, Chair CEN-CENELEC 21.
9. Voir le comité ISO/IEC JTC 1/SC42 Intelligence Artificielle.

AfIA
Association française
pour l'Intelligence Artificielle

# Plateforme de sensibilisation à la normalisation de l'Intelligence Artificielle

Par

**Louis MORILHAT**
*AFNOR*
*louis.morilhat@afnor.org*
*www.afnor.org*

**Caroline CHOPINAUT**
*Hub France IA*
*caroline.chopinaud@hub-franceia.fr*
*www.hub-franceia.fr*

**Chloé PLÉDEL**
*Hub France IA*
*chloe.pledel@hub-franceia.fr*
*www.hub-franceia.fr*

## Contexte

La question du fonctionnement sûr des logiciels est au cœur de nombreuses applications de tous les jours, qu'il s'agisse du transport (automobile, aviation, rail, etc.), des dispositifs de santé, des opérateurs d'intérêts vitaux. Les acteurs français se démarquent dans le domaine, en particulier grâce aux avancées produites par la recherche et industrialisées, entre autres dans les domaines de la conception/intégration de logiciel critique temps réel ou la preuve formelle de propriétés logicielles.

Le développement rapide des logiciels d'Intelligence Artificielle (IA) et leur diffusion dans tous les secteurs d'activité posent des questions spécifiques en termes de garanties sur leur « bon fonctionnement ». Que l'on pense à la sûreté d'une prise de décision « autonome » en temps réel, à des domaines ne tolérant pas l'erreur de décision (décisions de sécurité, de justice, diagnostic de santé, etc.) ou à des attentes d'équité de traitement qui exigent la garantie que les traitements ne sont pas biaisés, la confiance placée dans les systèmes intégrant de l'IA doit impérativement être développée comme ce fut le cas précédemment pour les logiciels déterministes « classiques ».

Pour ce faire, la Commission européenne ambitionne d'adopter une réglementation spécifique sur l'IA, l'*AI Act*, dont la première proposition de texte a été présentée le 21 avril 2021 [2]. Elle promeut une IA « sûre, fiable et innovante au bénéfice des entreprises et de toute la société européenne », soulignant « l'importance d'une approche équilibrée qui tient dûment compte de la protection de la sécurité et des droits fondamentaux, ainsi que les avantages économiques et sociétaux de la technologie » [1].

Ce projet de réglementation pose toutefois un véritable enjeu d'équilibre entre innovation, application, contrôle et gouvernance ; pour l'ensemble des acteurs industriels requérant à cette technologie sur le marché européen, et tout particulièrement pour l'écosystème des ETI/PME et des startups.

Afin de soutenir ce projet et permettre son adoption par tous les acteurs, la Commission européenne a décidé de s'appuyer sur des normes dites « harmonisées » qui établiront des déclinaisons techniques conformes au respect de la législation européenne.

Pour cela, la Commission européenne a fourni une requête de normalisation auprès de l'organisme européen dédié, le CEN-CENELEC, comportant dix points qui doivent être couverts par un corpus de normes techniques :

- *risk management system*
- *governance and quality of datasets*
- *record keeping*
- *transparency and information*
- *human oversight*
- *accuracy specifications*
- *robustness specifications*
- *cybersecurity specifications*
- *quality management system*
- *conformity assessment*

Par ce dispositif, la conformité aux normes harmonisées constituera une présomption de conformité à la réglementation européenne. Ces dernières concernent plus particulièrement les systèmes d'IA qualifiés à haut risque, qui effectueront une évaluation de conformité pour pouvoir obtenir le marquage européen dit « IA de Confiance », nécessaire à leur mise sur le marché européen.

## Le Grand Défi Intelligence Artificielle

Au regard des aspects technologiques et sociétaux de l'IA ainsi que du cadre réglementaire décrit ci-dessus, des actions ont été entreprises au niveau français.

Pour répondre aux enjeux dans des domaines considérés comme stratégiques, le gouvernement finance des Grands Défis destinés à favoriser l'émergence d'innovations de rupture et leur déploiement.

L'Intelligence Artificielle constitue l'un de ces Grands Défis, piloté par le Secrétariat Général Pour l'Investissement (SGPI), dont l'objectif fixé est de « sécuriser, certifier et fiabiliser les systèmes fondés sur l'Intelligence Artificielle ». Financé par le Programme d'investissement d'avenir (PIA) et le plan France Relance, il comprend 3 volets : technologique et applicatif d'une part, à travers le consortium Confiance.ai qui visent à développer des outils et méthodes pour une IA de confiance ; normatif d'autre part, en chargeant AFNOR de répondre aux enjeux suivants :

- définir une stratégie nationale pour la normalisation de l'IA, fédérer les acteurs industriels et académiques autour de cette ambition et mettre en application une feuille de route opérationnelle pour l'écosystème français,
- accroître la capacité d'influence pour favoriser la promotion des normes d'intérêt pour la France,
- structurer et animer la communauté d'acteurs (en favorisant notamment l'implication des PME, des ETI, des startups, de la recherche, etc.), en vue de produire des normes répondant aux besoins des acteurs français.

La feuille de route stratégique, publiée en mars 2022, identifie des axes prioritaires de normalisation visant notamment à répondre à cette requête de normes harmonisées de la part de la Commission européenne. Ainsi, des travaux de normalisation portés par la France sont en cours au niveau européen, avec l'élaboration de normes stratégiques, notamment les caractéristiques d'une IA de confiance et un catalogue des risques.

À ce stade, la sensibilisation et la mobilisation de l'écosystème français apparaît comme essentielle afin d'engager suffisamment de ressources pour défendre nos intérêts et s'assurer que les futures normes correspondent aux besoins des acteurs français et favorisent l'innovation.

## Sensibiliser les acteurs à la normalisation de l'IA

Dans cette perspective, AFNOR et le Hub France IA, avec le soutien de France Digitale,

ont souhaité se regrouper au sein d'un projet de plateforme de sensibilisation à la normalisation de l'IA.

Le projet de plateforme a été lancé en octobre 2022, pour une durée de 16 mois, avec la volonté de toucher un maximum d'entreprises et de les accompagner dans leur compréhension de la normalisation pour faire face aux enjeux de conformité de la réglementation, tout en poussant les acteurs à s'impliquer plus activement dans les projets de normes. Pour ce faire, le projet se donne les missions suivantes :

1. renforcer la maturité de l'écosystème IA, aussi bien utilisateurs que fournisseurs IA sur les enjeux de réglementation et de normalisation en IA ;
2. recueillir les besoins, attentes et ressentis de l'écosystème vis-à-vis des normes et des projets de réglementation ;
3. animer dans le temps et de façon pérenne la communauté pour l'accompagner dans la mise en place des exigences liées aux réglementations sur l'IA.

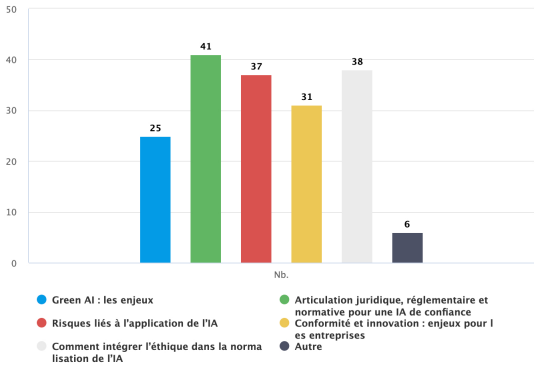
C'est au travers d'ateliers thématiques, de webinaires et d'un portail en ligne que le projet déroule ses actions en suivant ces trois grands axes.

## Les grandes phases du projet

Dans l'objectif de mobiliser l'écosystème pour assurer le suivi de la stratégie française et construire un réseau d'acteurs autour des enjeux de l'IA et de la normalisation, les premières actions du projet ont consisté à rassembler startups, PME et ETI lors d'une première réunion de travail qui a permis d'identifier les thèmes d'intérêts des participants. Une enquête a ensuite été lancée auprès d'une centaine de personnes intéressées par le projet pour recueillir leurs besoins et leurs attentes sur le sujet de la normalisation de l'IA.

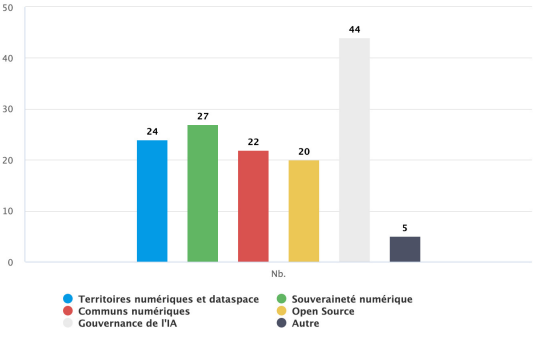
Les résultats issus de ces deux premières phases ont permis de mettre en évidence les thèmes prioritaires suivants pour des échanges avancés :

- la gestion des risques dans la mise en œuvre des systèmes d'IA (72% d'intérêt),
- comment intégrer l'éthique dans la normalisation de l'IA (74% d'intérêt),
- les articulations juridiques, règlementaires et normatives pour une IA de confiance (80% d'intérêt),
- conformité et innovation : enjeux pour les entreprises (60% d'intérêt)
- *Green AI* : les enjeux (49% d'intérêt)




Les participants ont aussi mis en évidence un besoin d'informations sur les thématiques suivantes :

- territoires numériques et *dataspace* (49% d'intérêt)
- souveraineté numérique (55% d'intérêt)
- communs numériques (45% d'intérêt)
- Open Source (40% d'intérêt)
- gouvernance de l'IA (80% d'intérêt)



Une fois ce recueil des besoins effectué, le projet s'organise en deux grandes phases qui se déroulent en parallèle :

- la mise en place d'ateliers thématiques permettant de construire un réseau d'acteurs autour des enjeux de la normalisation de l'IA,
- la mise en place d'actions de dissémination pour développer et partager une vision stratégique de la normalisation.

La phase d'ateliers vise principalement à :

- rassembler des parties prenantes diversifiées,
- fournir un espace d'échanges et de réflexion autour des enjeux de l'IA et favoriser les synergies,
- mettre en relation les différents cas d'usages pour identifier des éléments communs,
- permettre aux acteurs de faire remonter leurs besoins et retours d'expérience ?

Les ateliers portent sur les thèmes prioritaires identifiés dans la première phase du projet et s'articulent en sessions de 1h30 à 2h, d'une à trois sessions par thèmes en fonction des besoins des participants.

La phase de dissémination est soutenue par des actions de communication au travers de webinaires et de conférences, mais également par le biais du portail en ligne dont l'objectif est de maintenir un ensemble de contenus pertinents et pérennes de sensibilisation à la normalisation de l'IA.

Cette phase vise donc principalement à :

- sensibiliser les participants au futur cadre du marché européen de l'IA, à savoir l'articulation entre la législation, la normalisation et la certification sur l'IA,
- identifier les grands enjeux de la normalisation de l'IA et ses futurs impacts sur l'écosystème,
- promouvoir le rôle stratégique de la normalisation en matière de souveraineté, d'intelligence économique, de soutien à l'innovation ou encore de positionnement de marché.

## Assurer le suivi de la feuille de route française

Alors que les investissements pour l'innovation et l'accélération de l'industrialisation de l'IA sont nombreux, le cadre réglementaire et normatif se structure à court terme, avec des impacts majeurs pour le marché.

Aussi, ce projet s'inscrit à une étape charnière pour l'IA en France et en Europe, et l'implication des acteurs concernés nous semble indispensable.

Les résultats attendus de ce projet sont donc directement liés à une meilleure compréhension de la normalisation par la communauté des acteurs IA, quels que soient leur profil et leur niveau de maturité. Pour répondre aux problèmes rencontrés, prendre en considération les besoins métiers et ainsi apporter des solutions aux acteurs, notre ambition est de leur apporter des clefs de lecture quant aux travaux en cours et permettre leur participation à cette structuration du marché. Aussi, une des retombées principales du projet est d'être en mesure de mobiliser, sur le long terme, l'écosystème IA français, pour assurer le suivi de la feuille de route du Grand Défi IA par :

- le développement de la stratégie française de normalisation,
- la conduite de réflexions sur l'efficience de la participation française aux travaux de normalisation, en prenant en compte les priorités et les ressources disponibles,
- la facilitation des échanges avec la Commission de Normalisation IA, et son alimentation par des nouvelles perspectives,
- la promotion et l'accompagnement dans l'intégration d'acteurs intéressés dans des commissions de normalisation. Le développement et la mise à jour de la stratégie française de normalisation, au regard des ré-

flexions conduites et des besoins identifiés.

L'ensemble des résultats du projet seront disponibles très prochainement avec la mise en ligne d'un site dédié. Il est néanmoins possible de suivre les activités du projet via un groupe LinkedIn privé « Plateforme de Normalisation de l'IA ».

### Présentation des partenaires du projet

#### AFNOR

L'Association Française de Normalisation (Loi 1901) anime le système français de normalisation. Elle accompagne et guide les acteurs économiques pour élaborer les normes volontaires nationales, européennes et internationales. Près de 20 000 représentants d'entreprises, d'associations, de fédérations et de l'Etat participent chaque année à cet exercice de co-construction.

#### Hub France IA

Fondé en 2017, le Hub France IA est une association fédératrice de l'écosystème IA français avec pour ambition d'accélérer l'industrialisation et la croissance de l'IA au sein des entreprises, institutions ou collectivités. Regroupant plus d'une centaine d'adhérents, aussi bien Grands Groupes, PME/ETI, Ecoles, Institutions et Startups IA, le Hub France IA agit pour une IA appliquée et souveraine au service de projets opérationnels.

### Remerciements

Ce projet est réalisé au titre de l'action grands défis soutenue par le Fonds pour l'Innovation et l'Industrie.

### Références

[1] *Thierry Breton, Pour une Intelligence Artificielle fiable, sûre et innovante, Commission Européenne*, 2021.

[2] *Proposition de règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (dit « AI Act »)*, 21 avril 2021.

## Recherche sur les normes techniques pour l'IA et les droits fondamentaux à Télécom Paris

Par **Mélanie GORNET**
*I3 (UMR CNRS 9217)/ NOS – Numérique, Organisation et Société*
*IP Paris, Télécom Paris, Département Sciences Economiques et Sociales*
*melanie.gornet@telecom-paris.fr*

**Winston MAXWELL**
*winston.maxwell@telecom-paris.fr*

### Introduction

L'équipe *Operational AI Ethics* du laboratoire $i^3$ de Télécom Paris - Institut Polytechnique de Paris, mène des travaux interdisciplinaires autour de l'éthique de l'intelligence artificielle. Dans le cadre du projet ANR LIMPID sur les systèmes de reconnaissance d'images de confiance, nous nous sommes intéressés aux initiatives de normalisation de l'IA [17].

L'objectif est de comprendre comment le cadre législatif entourant les normes de sécurité des produits entrant sur le marché européen

peut s'adapter aux nouveaux systèmes d'IA. Les recherches actuelles portent notamment sur les normes relative à l'éthique de l'IA, particulièrement celles susceptibles de concerner les droits fondamentaux. Le rôle des normes et leur contribution à l'éthique de l'IA sont étudiés en regard des limites qu'elles présentent. Les processus de certification relative à l'IA « de confiance » sont également analysés, dans le but de discuter des protections effectives qu'ils apportent aux individus.

Une première partie de notre travail consiste en une cartographie des organismes de normalisation et de leurs activités relatives à l'éthique de l'IA. Dans une étude conjointe, nous examinons la notion algorithmique d'équité, sa transcription dans les propositions de normes et le décalage de ces définitions techniques avec le principe légal de non-discrimination.

## Dynamiques internationales

Le futur Règlement sur l'IA, ou *AI Act*, présenté en avril 2021 par la Commission Européenne [1], propose de s'appuyer sur une approche de normalisation des produits et des processus pour garantir la conformité des systèmes d'IA. Cette démarche permettrait, entre autres, de réduire les risques pour la sécurité et les droits fondamentaux des personnes concernées.

De nombreuses organisations se sont saisies du sujet de la normalisation de l'IA. Au niveau international, ces initiatives sont portées par un travail commun de l'*International Organization for Standardization* (ISO) et de la *International Electrotechnical Commission* (IEC), à travers le comité technique mixte ISO/IEC JTC 1. En son sein, le sous-comité (SC) 42 est chargé des normes relatives à l'intelligence artificielle. Les travaux d'autres sous-comités sont également pertinents, même s'ils se concentrent généralement sur un sous-domaine spécifique des technologies, comme l'ISO/IEC JTC 1/SC 37 sur la biométrie. Au niveau européen, les organismes nationaux de normalisation, comme l'Association Française de Normalisation (AFNOR) ou le *Deutsches Institut für Normung* (DIN) en Allemagne, collaborent au sein d'un comité technique conjoint (JTC) sur l'intelligence artificielle du Comité Européen de Normalisation (CEN) et du Comité Européen de Normalisation Électrotechnique (CENELEC) : le CEN-CLC/JTC 21. Parmi les thèmes identifiés, l'éthique occupe une place prépondérante [8]. L'Institut Européen des Normes de Télécommunication (ETSI) prépare également son lot de normes pour soutenir l'*AI Act* [21]. D'autres organismes travaillent sur des normes pour l'IA. C'est le cas notamment de la société savante *Institute of Electrical and Electronics Engineers* (IEEE) [4] ou encore du *National Institute of Standards and Technology* (NIST) [23] aux Etats-Unis. Enfin, certaines entreprises ou organismes indépendants développent leurs propres cadres d'évaluation, souvent dans le but d'aboutir à un label qui attesterait de l'éthique du produit d'IA ou de son constructeur.

## Différents types de normes

Parmi ces normes censées nous protéger contre la discrimination induite par les systèmes d'IA, nous distinguons différentes familles, selon les problèmes qu'elles traitent : elles peuvent être horizontales et générales, ou verticales et techno-spécifiques ; se centrer sur les produits de l'IA ou les processus qui mènent à leur fabrication ; contenir des taxonomies, des méthodes de conception ou des exigences de performance.

Au sein des différents types de normes, nous relevons notamment le développement récent des normes autour de la conception éthique. Elles ne sont pas spécifiques aux systèmes d'IA et se centrent généralement sur les

processus de développement - non sur les produits qui en résultent. Un bon exemple d'une norme de cette famille est la *IEEE Std 7000™-2021 : IEEE Standard Model Process for Addressing Ethical Concerns during System Design* [10].

Une autre famille de normes adresse la question des biais et de l'équité algorithmique. Elles visent à définir la notion de biais, lister les tests et mesures qui permettent d'évaluer si un système est biaisé, et dans la mesure du possible, de recenser les techniques permettant de les supprimer. *ISO/IEC TR 24027 :2021, Information technology – Artificial intelligence (AI) — Bias in AI systems and AI aided decision making* [19] est un exemple d'une telle norme.

Certaines de ces normes se centrent sur un domaine particulier. Par exemple, en biométrie, l'équité est abordée à travers les différentiels démographiques et les normes définissent des mesures et des procédures de tests adaptées à ces technologies précises. Les normes de cette catégorie sont toujours au stade de développement, comme l'*ISO/IEC WD 19795-10, Information technology – Biometric performance testing and reporting – Part 10 : Quantifying biometric system performance variation across demographic groups* [18]. Néanmoins, il est possible d'avoir une idée de ce qu'elles pourraient contenir en regardant les travaux du NIST comme le *Face Recognition Vendor Test (FRVT), Part 8 : Summarizing Demographic Differentials* [13].

## Rôle et limites des normes

Les normes sont des documents techniques destinés à établir des solutions communes à des exigences données [7]. Elles permettent notamment de définir un langage commun entre les acteurs, d'harmoniser les pratiques ou de définir le niveau de qualité des produits et des services [2]. Les normes peuvent remplir plusieurs rôles : elles aident les entreprises à optimiser les coûts et à augmenter leur efficacité [12], elles contribuent à dynamiser l'économie [16], à stimuler l'innovation [3], à encourager la concurrence entre les entreprises [5], et à assurer la protection des consommateurs [14].

Néanmoins, les normes font également face à de nombreux problèmes. Par exemple, les normes sur l'équité ou la réduction de biais dans les modèles d'apprentissage se heurtent inévitablement à la définition de l'équité algorithmique qui ne fait pas consensus [22]. Les différentes définitions sont parfois même contradictoires [20]. De plus, ces normes ne sauraient garantir l'absence totale de discrimination, elles ne font que définir des outils qui évitent de trop gros écarts de performance.

Certaines questions ne sont toujours pas traitées. Quels mécanismes mettre en place à l'échelle de l'entreprise pour encourager les bonnes pratiques de conception ? Quels tests globaux mettre en œuvre sur les systèmes entraînés pour réduire les risques de dysfonctionnement ? Quels dispositifs de contrôle *a posteriori* doivent être déployés ?

## Rôle et limites de la certification

Pour réguler l'IA, la Commission a fait le choix du modèle de la conformité [6]. Les normes deviennent ainsi un outil pour attester de cette conformité aux exigences essentielles définies dans l'*AI Act*, en particulier les normes européennes harmonisées qui bénéficient d'une présomption de conformité. Notamment, les systèmes jugés à « haut risque » devront être certifiés « Conformité Européenne » (CE). Ce type particulier de certification est obligatoire pour certains groupes de produits afin de pouvoir entrer sur le marché européen. Toutefois, le marquage CE est appliqué directement par le constructeur et ne signifie pas que l'UE a approuvé un produit comme étant sûr ou conforme [9].

En général, d'autres certifications pourront

s'appliquer à l'IA. Une fois une norme publiée, un organisme indépendant peut proposer un « certificat », agissant comme une garantie écrite et attestant que le produit ou service répond aux exigences définies dans la norme. Ainsi, la certification publique, dont fait partie le marquage CE, est destinée à faciliter le commerce tout en garantissant le respect des exigences réglementaires. La certification privée, quant à elle, consiste davantage en un repère de qualité pour le consommateur.

Mais ces outils de la mise en conformité peuvent également être critiqués. Le marquage CE a tendance à être, à tort, considéré par les consommateurs comme un gage de sécurité [11]. Pour les systèmes d'IA, il pourrait même être considéré comme une marque de protection des droits fondamentaux lorsque le produit a été testé selon des normes « éthiques ». Par exemple, le respect d'une norme sur l'équité algorithmique pourrait être considérée injustement comme une garantie de non-discrimination.

Au-delà d'une simple idée reçue de la part des consommateurs, cette apparente garantie pourrait rentrer en tension avec le travail de contrôle des juges, qui sont les seuls acteurs habilités à déterminer si un système est vraiment discriminatoire. Même si les normes européennes harmonisées sont souvent considérées comme une extension de la loi [24], elles n'ont pas vocation à modifier le niveau de protection des droits fondamentaux. Un système peut être conforme à une norme harmonisée, mais toujours être en violation des droits d'une personne.

## Analyse de risques et garantie des droits

L'*AI Act* est basé sur un système de classification des systèmes d'IA selon leur niveau de risque. Néanmoins, il persiste une tension entre cette approche « par les risques », proposée par la Commission, et l'approche axée sur le respect des droits fondamentaux, privilégiée par les tribunaux.

Si les normes peuvent permettre d'imaginer un banc de tests pour un critère de sûreté, à la manière des essais réalisés sur les jouets pour enfants avant leur mise sur le marché, il est plus difficile d'évaluer la non-discrimination de cette façon, car elle est contextuelle [25]. Ces normes définissent alors des critères techniques, censés demeurer neutres, alors qu'ils se basent sur des jugements de valeurs moraux et culturels.

De même, l'écart entre les définitions mathématiques de l'équité algorithmique et celle légale du droit à la non-discrimination créent différents niveaux de lecture des normes. Ainsi, les systèmes d'IA ne pourront jamais être parfaitement équitables et il sera donc impossible de garantir l'absence de discrimination sous tous les angles. Des choix seront nécessaires pour définir quels types et quels niveaux d'erreurs sont satisfaisants, et donc quelle discrimination est acceptable. Cette démarche est courante en analyse de risques. Ainsi, pour la sûreté d'une centrale nucléaire, le risque d'un accident sera quantifié, et la centrale pourra ouvrir si le risque est en deçà d'un certain seuil d'acceptabilité [15]. Toutefois, dans le cadre de la non-discrimination, définir un tel seuil dérange. En effet, si cette réflexion peut être menée au niveau statistique, elle rentre en conflit avec l'approche des tribunaux qui reconnaissent aux droits des personnes un caractère absolu. Une discrimination restera condamnable même si elle est statistiquement rare.

Les rôles des normes et de la certification sont donc à repenser : non comme des marques de garantie des droits mais comme des outils réduisant les pratiques dangereuses. Ils doivent alors être pensés en parallèle, et non en remplacement, d'un régime solide de responsabilité.

## Références

[1] Proposition de règlement du parlement européen et du conseil établissant des règles harmonisées concernant l'intelligence artificielle (dit « AI Act »), 21 avril 2021.

[2] AFNOR. Parler normes couramment. L'essentiel, 2014.

[3] R. H. Allen and R. D. Sriram. The Role of Standards in Innovation. *Technological Forecasting and Social Change*, 64(2) :171–181, 2000.

[4] IEEE Standards Association. AIS Standards. IEEE portfolio of AIS technology and impact standards and standards projects.

[5] K. Blind. The impact of standardisation and standards on innovation. In *Handbook of Innovation Policy Impact*, pages 423–449. 2016.

[6] C. Castets-Renard and P. Besse. Ex ante Accountability of the AI Act : Between Certification and Standardization, in Pursuit of Fundamental Rights in the Country of Compliance, 2022.

[7] T. Cellucci. Developing Operational Requirements - A Guide to the Cost-Effective and Efficient Communication of Needs, 2008.

[8] CEN-CENELEC. CEN-CENELEC Focus Group Report : Road Map on Artificial Intelligence (AI), 2020.

[9] European Commission. Commission Notice - The 'Blue Guide' on the implementation of EU product rules 2022. Technical report, Information from European Union Institutions, Bodies, Offices and Agencies, 2022.

[10] Systems and Software Engineering Standards Committee. IEEE Std 7000™-2021 : IEEE Standard Model Process for Addressing Ethical Concerns during System Design. 2021.

[11] A. de Tervueren. ANEC Position Paper on CE marking "Caveat Emptor - Buyer Beware", 2012.

[12] H. J. de Vries. *Standardization : A Business Approach to the Role of National Standardization Organizations*. Springer US, Boston, MA, 1999.

[13] D. L Duewer. Face Recognition Vendor Test (FRVT) Part 8 : Summarizing Demographic Differentials, 2022.

[14] Union européenne. Les normes en Europe, Your europe.

[15] B. Fischhoff. "Acceptable Risk" : The Case of Nuclear Power. *Journal of Policy Analysis and Management*, 2(4) :559–575, 1983.

[16] DIN German Institute for Standardization e. V. *Economic benefits of standardization - Summary of results*. Beuth Verlag, DE, 2000.

[17] M. Gornet and W. Maxwell. Intelligence artificielle: normes techniques et droits fondamentaux, un mélange risqué. *The Conversation*, 2022.

[18] ISO/IEC. ISO/IEC WD 19795-10, Information technology — Biometric performance testing and reporting — Part 10 : Quantifying biometric system performance variation across demographic groups, à paraître.

[19] ISO/IEC. ISO/IEC TR 24027 :2021, Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making, 2021.

[20] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv :1609.05807*, 2016.

[21] M. Mueck and *et alii*. ETSI White Paper No. #52. ETSI Activities in the field

of Artificial Intelligence Preparing the implementation of the European AI Act, 1st Edition. December 2022.

[22] A. Narayanan. 21 fairness definitions and their politics, 2019.

[23] NIST. NIST AI Program. Artificial Intelligence : The Vitals, 2023.

[24] A. van Waeyenberge and D. Restrepo. James Elliot construction : A "New(ish) approach" to judicial review of standardisation. *European Law Review*, 42 :882–893, 2017.

[25] S. Wachter, B. Mittelstadt, and C. Russel. Why fairness cannot be automated : Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, 41, 2021.

## Les trois degrés de transparence des décisions individuelles fondées sur l'utilisation de systèmes d'IA – Du drapeau rouge de Turing à l'obligation d'explicabilité

Par **Pascal ALIX**
*Avocat associé/DPO Externe – Lead auditor SGS*
*Virtualegis AARPI*

L'intelligence artificielle, telle que définie par la proposition de règlement sur l'intelligence artificielle [3] (ci-après « *AI Act* »), constitue l'aboutissement actuel des techniques informatiques. Or, l'article premier de loi « informatique et libertés » [2] proclame, sur le modèle des grandes déclarations des droits de l'Homme : « L'informatique doit être au service de chaque citoyen [. . . ] Elle ne doit porter atteinte ni à l'identité humaine, ni aux droits de l'homme, ni à la vie privée, ni aux libertés individuelles ou publiques. » Ces principes clairs, anciens, revêtus de la force normative attachée aux lois d'ordre public [10], ne sont pas remis en cause par les autres textes nationaux, ni par ceux en préparation au niveau de l'Union européenne.

Or, comment s'assurer que la mise en œuvre d'un système d'intelligence artificielle (système d'IA) ne porte atteinte ni à l'identité humaine, ni aux droits de l'homme, ni à la vie privée, ni aux libertés individuelles ou publiques ? Comment, face à l'opacité de certains algorithmes d'apprentissage automatique, permettre aux utilisateurs finaux, aux personnes concernées, aux actuels et futurs régulateurs et, le cas échéant, aux institutions judiciaires, d'effectuer un contrôle du respect de ces principes ?

Un sondage réalisé en France, en Allemagne et dans les Pays-Bas en novembre 2022 [4] révèle notamment que la plupart des Français interrogés estiment important d'être informés sur l'usage des systèmes d'IA, notamment lorsqu'il s'agit d'évaluer ou de faire une prédiction sur leur comportement, qu'ils « estiment décisif d'avoir accès à une explication de la logique du modèle, dans une langue simple » [17] et que la majorité d'entre eux souhaite pouvoir s'adresser à un humain, notamment en cas de décision

10. Auxquelles on ne peut déroger « par des conventions particulières », c'est-à-dire par des dispositions contractuelles (article 6 du code civil).
11. Par ex. en cas de procédure de recrutement avec sélection de dossiers assistée par des outils algorithmiques.

négative [11]. De manière générale, la transparence des systèmes d'IA est un enjeu important pour la confiance et l'acceptabilité de ces technologies, en particulier dans des secteurs d'activités « critiques » tels que les processus démocratiques, la santé ou la justice.

Le Conseil d'État, dans son étude de 2022 sur l'intelligence artificielle [13], propose notamment un principe fondamental de transparence [12], sans l'existence duquel les droits des personnes concernées par l'utilisation d'un système d'IA ne peuvent être exercés. Dans son « projet révisé de convention [cadre] sur l'intelligence artificielle, les droits de l'homme, la démocratie et l'État de droit » [11], le Conseil de l'Europe propose un « principe de transparence et de contrôle » selon lequel les États-membres du Conseil de l'Europe intégreraient dans leur droit interne « des mécanismes de contrôle ainsi que des exigences de transparence et d'auditabilité appropriées . . . »

La transparence conduit à exiger des informations diverses sur les systèmes d'IA dans le cadre de différents cas d'usages (selon les organismes, les secteurs, les techniques utilisées et leurs finalités).

Lorsqu'une décision individuelle est fondée sur l'utilisation d'un système d'IA, la première des exigences consiste à informer immédiatement les personnes physiques qu'elles interagissent avec un système d'IA ou qu'un système d'IA traite des données à caractère personnel qui les concernent (1). La deuxième consiste à fournir des informations sur les caractéristiques du système d'IA et les conditions et modalités de sa mise en œuvre (2). La troisième consiste à exiger des concepteurs, fournisseurs et utilisateurs professionnels qu'ils fournissent les moyens d'exercer un contrôle effectif sur le fonctionnement du système d'IA et sur les moyens techniques et logiques utilisés pour parvenir au résultat (explicabilité) (3).

## Le premier degré de transparence : l'obligation d'alerte

Lorsqu'une entreprise ou une administration entreprend de « déléguer des tâches à des systèmes automatiques » [10] pour prendre des décisions individuelles, les personnes concernées s'attendent généralement, nous l'avons vu, à en être informées. L'information initiale sur l'utilisation d'un système d'IA constitue, pour certains auteurs, l'un des éléments de « la transparence des algorithmes » [5].

### Le Turing's red flag

Dans le *Locomotive Act 1865* du Royaume Uni, également dénommé « *Red Flag Act* » il a notamment été prévu que les voitures à moteur thermique empruntant les voies publiques – et dont la vitesse était limitée à 6 km/h sur les routes et 3 km/h dans les villes – devaient être précédées d'un porteur d'un drapeau rouge pour avertir les autres usagers de la route [19]. L'amendement de 1878 ayant augmenté la vitesse limite à 23 km/h, d'autres techniques d'avertissement ont dû être utilisées, comme l'avertisseur sonore [13]. Dans un article intitulé « Turing's Red Flag » [18], Toby Walsh, professeur d'intelligence artificielle à l'université de New South Wales, s'en inspirant, a émis l'idée d'une « loi du drapeau rouge de Turing »» selon laquelle un « système autonome » devrait être « conçu de manière à ne pas être confondu avec autre chose qu'un système autonome », et « s'identifier au début de toute interaction avec un autre agent [14]. »

12. Comportant « à tout le moins, le droit d'accès à la documentation du système, une exigence de loyauté consistant à informer les personnes de l'utilisation d'un SIA à leur égard, l'auditabilité du système par les autorités compétentes ainsi que la garantie d'explicabilité. »
13. La marque « Klaxon » de l'avertisseur sonore devenu électrique a été déposée en 1908.
14. Autrement dit, un être humain.

Le droit français exige une transparence renforcée de la part des administrations publiques. Le principe de la Déclaration des droits de l'homme et du citoyen selon lequel « La société a le droit de demander compte à tout agent public de son administration » (article 15) est précisé par le code des relations entre le public et l'administration (CRPA, articles L. 300-1 à L. 351-1.). Il s'applique notamment aux décisions fondées, exclusivement ou non, sur le résultat généré par un système d'IA, qui constituent une catégorie des « décisions algorithmiques ». Or, l'article L. 311-3-1 du CRPA prévoit qu'en principe « une décision individuelle prise sur le fondement d'un traitement algorithmique comporte une mention explicite en informant l'intéressé ».

Cette règle s'applique à toute décision fondée sur le résultat généré par un système d'IA, sauf si son application est susceptible de porter atteinte, notamment, au secret de la défense nationale, à la sûreté de l'État, à la sécurité publique, à la sécurité des personnes ou à la sécurité des systèmes d'information des administrations, à la monnaie et au crédit public, au déroulement de procédures judiciaires et « à la recherche et à la prévention, par les services compétents, d'infractions de toute nature ou [. . . ] aux autres secrets protégés par la loi. »

Le professionnel de santé qui utilise un dispositif médical « comportant un traitement de données algorithmique dont l'apprentissage a été réalisé à partir de données massives » est, par ailleurs, tenu de s'assurer que la personne concernée en a été informée (article L. 4001-3, I du Code la santé publique).

L'*AI Act* retient le principe d'une information des personnes physiques concernées « qu'elles interagissent avec un système d'IA » « de manière claire et reconnaissable au plus tard au moment de la première interaction ou de la première exposition » (article 52 et considérant 70) [3]. Cette obligation s'appliquerait en présence d'un « système d'IA destiné à interagir avec des personnes physiques ou à générer du contenu », dans la mesure où un tel système peut présenter « des risques spécifiques d'usurpation d'identité ou de tromperie ». Un tel devoir d'alerte devrait également s'appliquer à la reconnaissance biométrique, à la reconnaissance des émotions, à la manipulation des images ou contenus audio ou vidéo présentant une ressemblance avec des personnes, des objets, de lieux ou d'autres entités ou évènements existants (hypertrucage ou « deepfake »).

La Commission Nationale Consultative des Droits de l'Homme (CNCDH) recommande d'étendre le champ d'application de cette obligation, qui devrait concerner, selon elle, non seulement les systèmes d'IA destinés à interagir avec les personnes physiques [15], mais dès que les personnes physiques font l'objet d'une décision fondée en tout ou en partie sur un « traitement algorithmique » [7], notion qui peut être comprise comme traitement de données à caractère personnel mettant en œuvre un système d'IA. De manière générale, la CNCDH estime que l'information relative à l'intervention d'un système d'IA dans la décision « est un préalable qui ne doit souffrir aucune exception » [7].

### L'exception

Selon l'*AI Act* [3], l'utilisateur professionnel du système d'IA pourrait être exempté de son devoir d'information initiale s'il « ressort clairement du point de vue d'une personne physique normalement informée et raisonnablement attentive et avisée, compte tenu des circonstances et du contexte d'utilisation » qu'elle interagit avec un système d'IA (considérant 70).

15. Tels que les agents conversationnels (« chatbots », par exemple, ChatGPT).

## Le deuxième degré de transparence : les obligations d'information

Toute décision individuelle résultant, en tout ou en partie, de l'utilisation d'un système d'IA suppose, en outre, la fourniture d'informations sur différents éléments du « traitement algorithmique » sur lequel elle repose, sur ses finalités exactes, sur le résultat du traitement des données, ainsi que sur les moyens d'y parvenir (caractéristiques des algorithmes, logique sous-jacente, données traitées, etc.)

### Les informations à fournir par tous les organismes utilisateurs de systèmes d'IA

Dès lors que le système d'IA traite des données à caractère personnel, l'organisme est notamment soumis en France à la loi informatique et libertés et au RGPD, desquels il résulte qu'il :

- doit fournir à la personne concernée toutes les informations exigées par l'article 13 ou 14 du RGPD (finalités, bases légales, « informations utiles concernant la logique sous-jacente », etc.), dans les conditions prévues par l'article 12 du même règlement (communication d'informations « d'une façon concise, transparente, compréhensible et aisément accessible, en des termes clairs et simples ») ;
- doit se conformer à l'article 22 du RGPD, qui prévoit la possibilité de permettre aux personnes concernées de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement algorithmique automatisé « produisant des effets juridiques la concernant ou l'affectant de manière significative de façon similaire » sauf à être en mesure de démontrer que la décision algorithmique est, dans le cas considéré, expressément autorisée par le droit français ;
- doit mettre en œuvre des mesures techniques et organisationnelles appropriées « pour s'assurer et être en mesure de démontrer que le traitement est effectué conformément au présent règlement » (« *accountability* » – article 24 du RGPD).

À ces obligations vont s'ajouter, à horizon 2025, celles issues de l'*AI Act*. Or, si l'utilisation professionnelle d'un système d'IA « à haut risque [16] » impose, en l'état actuel de l'*AI Act*, le respect des notices d'utilisation, le contrôle humain ainsi que le contrôle et l'enregistrement de données, et si un devoir d'alerte existe, nous l'avons vu, pour certains systèmes d'IA, il ne prévoit, à la charge des utilisateurs professionnels, aucune obligation générale d'information des personnes concernées.

Les articles 9 (*Système de gestion des risques*), 11 (*Documentation technique*), 12 (*Enregistrement*) et 13 (*Transparence et fourniture d'informations aux utilisateurs*) et 20 (*Journaux générés automatiquement*) et 29.5 (*Obligations des utilisateurs de systèmes d'IA à haut risque*) de l'*AI Act* prévoient des obligations, complétées par les annexes IV (*Documentation technique visée à l'article 11, paragraphe 1*) et VII (*Conformité fondée sur l'évaluation du système de gestion de la qualité et l'évaluation de la documentation technique*) permettant, en revanche, une « *accountability* », dans des conditions à déterminer ultérieurement, notamment par le droit national et les règles d'interprétation établies notamment par l'« autorité nationale compétente [17] ».

### Les informations à fournir par l'administration publique qui utilise un système d'IA (la transparence administrative)

En sus des obligations communes précitées, les administrations publiques qui procèdent à

16. Selon la classification de l'annexe III.
17. Qui pourrait, en France, être la CNIL, selon la recommandation du Conseil d'État [13].

un « traitement algorithmique » (automatisé) sont tenues, en premier lieu, de mentionner, dans le cadre d'une « mention explicite » obligatoire (article R. 311-3-1-1 du CRPA) :

- la finalité poursuivie par le traitement algorithmique,
- le droit d'obtenir la communication des règles définissant le traitement et des principales caractéristiques de sa mise en œuvre,
- les modalités d'exercice de ce droit à communication et de saisine, le cas échéant, de la CADA [18].

En pratique, la mention explicite doit figurer en ligne et sur les documents (avis, notifications) qui procèdent à la notification de la décision à la personne concernée. Cette mention doit être visible par les personnes concernées [15].

Les administrations publiques sont tenues, en second lieu, de communiquer à l'intéressé qui en fait la demande « les principales caractéristiques » de la mise en œuvre du « traitement algorithmique » [19]. Celles-ci comportent les informations suivantes (« sous une forme intelligible ») :

- le degré et le mode de contribution du traitement algorithmique à la prise de décision ;
- les données traitées et leurs sources ;
- les paramètres de traitement et, le cas échéant, leur pondération, appliqués à la situation de l'intéressé ;
- les opérations effectuées par le traitement.

La loi informatique et libertés [2] précise que pour les décisions administratives algorithmiques, l'administration « s'assure de la maîtrise du traitement algorithmique et de ses évolutions afin de pouvoir expliquer, en détail et sous une forme intelligible, à la personne concernée la manière dont le traitement a été mis en œuvre à son égard » (article 47, alinéa 2).

Dans son avis sur le projet de loi pour une République Numérique [6], la CADA a retenu que « pour présenter un effet utile, les dispositions du nouvel alinéa ajouté à l'article L. 311-3 [. . . ] doivent être comprises comme ouvrant en outre à ces personnes le droit d'obtenir de l'administration, en complément de la communication éventuelle du code source, dont la compréhension nécessite des compétences techniques en codage informatique, des explications complémentaires, explicitant les règles de traitement mises en œuvre et les principales caractéristiques de celle-ci ».

Le département ETALAB de la direction interministérielle du numérique (DINUM) déduit, quant à lui, des exigences de transparence et d'intelligibilité que les traitements de machine learning qui ne peuvent pas être expliqués peuvent être utilisés uniquement comme des outils d'aide à la décision, en gardant un humain dans la boucle [14].

Les informations qui devront être fournies par les administrations publiques dépendront de la catégorie du système d'IA utilisé [20]. Les autorités, agences ou organismes publics doivent fournir un certain nombre d'informations lors de leur enregistrement en tant qu'opérateurs [21], parmi lesquelles des informations techniques obligatoires (description de la destination du système d'IA, notice d'utilisation en format électronique).

### Les informations à fournir par les professionnels de santé

Un nouvel article L. 4001-3 a été inséré dans le code de la santé publique en 2021 [1] au sujet des dispositifs médicaux « comportant

18. Commission d'accès aux documents administratifs.
19. Article L.311-3-1 du CRPA.
20. Risque minime, risque faible, haut risque ou risque critique justifiant une interdiction.
21. Annexe VIII.

un traitement de données algorithmique dont l'apprentissage a été réalisé à partir de données massives ». Désormais, le professionnel de santé qui utilise le dispositif médical précédemment décrit s'assure que non seulement la personne concernée en a été informée, mais également qu'elle est « le cas échéant, avertie de l'interprétation qui en résulte », ce qui fait reposer sur le professionnel de santé de déterminer les conditions et modalités de l'information sur le résultat. En revanche, l'information sur les moyens techniques pour parvenir au résultat n'est pas prévue par le code de la santé publique.

### Les limites de l'obligation légale d'information

Sans nier l'utilité de la « transparence des systèmes d'IA » [22], renforcée lorsque ces systèmes sont utilisés par des organismes publics pour prendre des décisions individuelles, certains auteurs [16] considèrent qu'il ne s'agit que d'une « communicabilité des documents explicatifs des algorithmes » qui « n'améliorent pas la transparence de l'algorithme en lui-même. »

## Le troisième degré de transparence : l'obligation d'explicabilité

La notion d'explicabilité n'est pas présente dans le droit actuellement en vigueur en France. Selon la CNIL, qui recommande de mettre en place des mesures permettant l'explicabilité des systèmes d'IA, l'explicabilité est « la capacité de mettre en relation et de rendre compréhensible les éléments pris en compte par le système d'IA pour la production d'un résultat. » [8].

Elle l'est, en revanche, de manière implicite dans l'*AI Act*, en ce qu'elle exige pour les systèmes d'IA à haut risque la fourniture d'un certain nombre d'informations techniques.

L'annexe IV de l'*AI Act* décrit ce que doit contenir « la documentation technique visée à l'article 11, paragraphe 1 », à savoir notamment :

- une description générale du système,
- une description détaillée des éléments du système d'IA et du processus de développement et notamment « la logique générale du système d'IA et des algorithmes ; les principaux choix de conception, y compris le raisonnement et les hypothèses retenues », « la description de l'architecture du système expliquant la manière dont les composants logiciels s'utilisent et s'alimentent les uns les autres ou s'intègrent dans le traitement global ; les ressources informatiques utilisées pour développer, entraîner, mettre à l'essai et valider le système d'IA » et « le cas échéant, les exigences relatives aux données en ce qui concerne les fiches décrivant les méthodes et techniques d'entraînement et les jeux de données d'entraînement utilisés. . . », ainsi que « les procédures de validation et de test utilisées ».

Mais les obligations et mécanismes destinés à assurer une telle transparence des systèmes d'IA ne concernent que les systèmes d'IA à haut risque. De sorte que la confiance des personnes concernées dans les systèmes d'IA qui ne feront pas partie de cette catégorie ne sera pas nécessairement garantie [9]. Tel sera le cas en particulier des systèmes d'IA utilisés par des organismes privés dont l'obligation de transparence se résumera, en droit interne, à la nécessité de respecter les dispositions du RGPD lorsque le système d'IA traitera des données à caractère personnel non anonymisées. Cepen-

22. Présentée dans l'*AI Act* comme destinée à « assurer le respect des obligations pertinentes incombant à l'utilisateur et au fournisseur » (article 13) et à « remédier à l'opacité qui peut rendre certains systèmes d'IA incompréhensibles ou trop complexes » (considérant 47).

dant, selon les régulateurs [12], les obligations d'information, de transparence et d'explicabilité des algorithmes « devraient concerner les algorithmes du secteur privé comme ceux du secteur public ».

Quoi qu'il en soit les systèmes d'IA ayant des conséquences sur les droits des personnes concernées ne pourront être acceptés durablement que si les concepteurs et/ou les développeurs mettent en place des mesures permettant leur explicabilité,

- soit en privilégiant des algorithmes interprétables [23], ce qui est fréquemment requis de facto dans certains secteurs fortement réglementés, comme la banque ou l'assurance, lorsque l'utilisateur doit démontrer au régulateur les conditions des prises de décision,
- soit en ajoutant une surcouche d'explicabilité [24].

Les algorithmes interprétables étant souvent moins performants que les algorithmes de *deep learning* et les surcouches d'explicabilité ayant un coût et n'offrant qu'une garantie de véracité limitée, le choix dépend du système et du contexte de son utilisation.

## Références

[1] *Loi n° 2021-1017 du 2 août 2021 révisant la loi de bioéthique*.

[2] *Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés*.

[3] Proposition de règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (dit « AI Act »), COM/2021/206 final, 21 avril 2021. Voir la version actuelle de l'*AI Act* : Orientation générale du conseil de l'Union européenne modifiant la proposition d'*AI Act*, document ST 15698 2022 INIT, 6 décembre 2022.

[4] Sondage réalisé par Research Collective, institut de sondage polonais, pour la Fondation Panoptykon, ONG de défense des droits numériques, novembre 2022.

[5] B. Barraud. L'algorithmisation de l'administration. *Revue Lamy Droit de l'immatériel*, 150 :42–54, 2018.

[6] CADA. *Avis sur le projet de loi pour une République Numérique*, 19 novembre 2015, N° 20155079.

[7] CNCDH. Avis relatif à l'impact de l'intelligence artificielle sur les droits fondamentaux. *17e recommandation*, pages 27–28, 7 avril 2022.

[8] CNIL. Fiche définition « explicabilité ».

[9] CNIL. Utiliser un système d'IA en production.

[10] CNIL. Comment permettre à l'Homme de garder la main ? Rapport sur les enjeux éthiques des algorithmes et de l'intelligence artificielle, p. 26, décembre 2017.

[11] Comité sur l'intelligence artificielle (CAI) Conseil de l'Europe. Projet zero révisé de convention [cadre] sur l'intelligence artificielle, les droits de l'homme, la démocratie et l'Etat de droit – CAI(2023)01_FR, 6 janvier 2023.

[12] Défendeur de droits et CNIL. Étude Algorithmes : prévenir l'automatisation des discriminations, Mai 2020.

23. Comme les régressions linéaires, les arbres de décision, les systèmes à base de règles de type *RuleFit* ou les classifications basées sur les réseaux bayésiens.

24. Tels LIME, qui permet d'expliquer un résultat en approximant par une fonction « simple » la fonction de décision du système au voisinage du résultat, le *Partial Dependance Plot* (PDP), l'*Individual Conditional Expectation* (ICE), qui délivrent une explication visuelle au comportement du modèle ou l'*Adversarial Example*, dont l'objectif est de permettre la compréhension de ce que le modèle a accompli au moyen de quelques exemples.

[13] Conseil d'État (Étude à la demande du Premier ministre). Intelligence artificielle et action publique : construire la confiance, servir la performance, 31 mars 2022.

[14] ETALAB. Les algorithmes publics : enjeux et obligations.

[15] ETALAB. Fiche pratique : l'obligation de mention explicite.

[16] E. Mouriesse. L'opacité des algorithmes et la transparence administrative. *Revue française de droit administratif*, page 45, 2019.

[17] M. Saliou. Les Français veulent des explications claires sur les modèles algorithmiques. Je peux parler à un humain ?, 22 février 2023.

[18] T. Walsh. Turing's red flag – a proposal for a law to prevent artificial intelligence systems from being mistaken for humans. *Communications of the ACM*, 59(7) :34–37, 2016.

[19] Wikipedia. Locomotive Act.

# L'injonction à la transparence : un levier réglementaire à double tranchant pour les organisations

Par

**Louis VUARIN**
*I3 SES / Telecom Paris*
*Institut Polytechnique de Paris*
*louis.vuarin@telecom-paris.fr*
*louis.vuarin@polytechnique.edu*

**Véronique STEYER**
*I3 CRG / École Polytechnique*
*Institut Polytechnique de Paris*
*veronique.steyer@polytechnique.edu*

## Introduction

Réguler l'IA en exigeant davantage de transparence : un principe de plus en plus populaire, et qui semble se concrétiser progressivement au sein notamment de l'*AI Act* [1]. Pour autant, de telles injonctions à la transparence produisent-elles les effets attendus dans les organisations ? En particulier, la transparence est-elle vraiment un levier performant en matière de gestion des risques ? L'abondante littérature en sciences de gestion, et particulièrement en théorie des organisations appliquée à la gestion des risques, apporte un regard plutôt mitigé sur la question. Ces travaux alertent sur la possibilité de voir émerger une réglementation générant des dynamiques organisationnelles potentiellement contreproductives au regard des attentes affichées.

## L'exigence de transparence : une panacée ?

De nombreux arguments plaident en faveur d'une exigence accrue de transparence pour les organisations qui conçoivent, commercialisent ou utilisent des IA. Pour lutter contre les biais et discriminations (au titre des origines socio-ethniques supposées, de critères géographiques, patrimoniaux, de genre, de préférences

sexuelles, politiques, etc.) qui inquiètent la communauté académique et alimentent le débat sur l'éthique de l'IA [26, 30, 36, 51], la transparence offrirait une forme de garantie de vérification et de correction des discriminations non souhaitées, qu'elles résultent des algorithmes ou des bases de données à partir desquelles ils sont entraînés. Sur le long terme, en accentuant les capacités de contrôle par les développeurs eux-mêmes ainsi que par des tiers externes [31, 34], la transparence renforcerait la confiance entre parties prenantes [17, 55]. En théorie, ce droit de regard, potentiellement renforcé par un protocole délibératif adéquat [10], provoquerait un cercle vertueux permettant de faire dialoguer et même converger utilisateurs et concepteurs, contribuant *in fine* à l'acceptabilité sociétale de cette technologie [13, 48, 49]. Fort de cet argumentaire, l'exigence de transparence s'est progressivement imposée comme un incontournable des chartes éthiques pour les organisations et méta-organisations qui se sont lancées dans ce type de démarche [20, 23, 25]. Dans leur méta-analyse de 84 chartes éthiques, Jobin *et al.* [23] identifiaient ainsi la transparence comme le principe le plus fréquemment mis en avant (73/84 documents), devant notamment la justice et l'équité (*justice & fairness*, 68/84), l'innocuité (*non-maleficence*, 60/84), la responsabilité (*responsibility*, 60/84) ou encore le respect de la vie privée (*privacy*, 47/84). Dans la même veine, la législation européenne reprend aussi ce principe, en l'intégrant graduellement comme une exigence réglementaire pour les organisations qui se lancent dans la conception ou la commercialisation d'IA. Ainsi, l'*AI Act*, dont une première mouture fut révélée par la Commission européenne en avril 2021[1], accorde une importance significative au principe de transparence [46]. Dans ses dispositions générales (Titre I), l'*AI Act* se fixe ainsi comme objet d'établir « des règles harmonisées en matière de transparence applicables aux systèmes d'IA destinés à interagir avec des personnes physiques, aux systèmes de reconnaissance des émotions et de catégorisation biométrique, et aux systèmes d'IA utilisés pour générer ou manipuler des images ou des contenus audio ou vidéo » (article 1.d). L'article 13 précise les règles de « transparence et fourniture d'informations aux utilisateurs » des IA dites « à haut risque », et l'article 52 détaille les « obligations de transparence pour certains systèmes d'IA », par exemple pour informer les utilisateurs lorsqu'ils interagissent avec des IA (article 52.1), pour des IA utilisées dans la reconnaissance des émotions ou la catégorisation biométrique (article 52.2), ou encore pour encadrer l'utilisation d'IA à des fins de manipulation d'images, de sons et de vidéos (article 52.3) [25]

Mais la transparence représente-t-elle vraiment une panacée face aux nombreux défis éthiques de l'IA ? Le succès du concept de transparence ne doit pas cacher certains raccourcis éthiques qui alimentent son argumentaire, et de nombreux auteurs mettent en garde envers les insuffisances théoriques de ce concept populaire [6, 7, 18, 20, 23, 25, 50]. Ces analyses développent deux lignes de critiques. En premier lieu, il s'agit de souligner le caractère imprécis du concept, qui tire son succès d'une métaphore facilement compréhensible à première vue, mais dont la définition est instable selon les acteurs qui la formulent [7, 25]. À cause de ce flou, les modalités d'application restent souvent à géométrie variable, peu contraignantes dans la pratique, avec dans certains cas le risque d'une stratégie d'affichage sans réel effet assimilable à une forme d' « ethic washing » [8, 32]. Par ailleurs, l'idéal de transparence se retrouve de facto limité par la tech-

25. Comme les « *deep fakes* », dont la création est de plus en plus facile et perfectionnée grâce à l'IA [5] mais dont l'encadrement par la loi pose de réelles questions juridiques et philosophiques [45].

nique [16, 26] : la plupart des algorithmes qui alimentent le boom actuel de l'IA sont structurellement opaques. Dès lors qu'ils ont été entraînés, ces algorithmes deviennent des « boites noires » (*black boxes*) trop complexes pour être analysés dans le détail, sauf au prix d'efforts importants et avec une marge interprétative significative.

Mais au-delà des insuffisances du concept, il s'agit aussi d'explorer les effets secondaires de la transparence sur les organisations qui devront en appliquer les exigences. Car, si un idéal restant inefficace dans la pratique est dommageable pour ceux qui y investissaient quelque forme d'espoir, il devient préoccupant pour la société dans son ensemble lorsqu'il s'avère produire des effets contre-intuitifs voire contre-productifs sur les acteurs et les organisations qu'il prétendait améliorer. Pour cela, il s'agit d'étudier l'impact organisationnel de l'injonction de transparence, et notamment d'anticiper les diverses dynamiques de jeux de pouvoir et de distorsions de l'attention qu'il peut produire. En effet, les sciences de gestion, notamment les auteurs qui se sont concentrés sur l'étude de la gestion des risques, ont montré que la transparence peut provoquer des conséquences inattendues sur le comportement des acteurs [2, 3, 4, 22, 38, 39, 40], obligeant à fortement relativiser les bénéfices éthiques attendus.

## Mieux anticiper les effets de la transparence sur le comportement des acteurs dans les organisations

Dans cette perspective, nous proposons de mettre en valeur les deux faces de la médaille. Dans un premier temps, nous listons les dynamiques organisationnelles vertueuses qui sont suggérées par l'idéal de transparence appliqué aux organisations concevant ou commercialisant des IA. Puis, nous proposons de nous appuyer sur la littérature des sciences de gestion pour mieux caractériser les effets secondaires qui pourraient résulter de l'application des injonctions à la transparence, telle que notamment formulée dans l'*AI Act*[26].

### Côté pile : amorcer des dynamiques organisationnelles vertueuses

À première vue, l'*AI Act* semble élaboré selon une logique directement influencée par un objectif de diminution de l'ampleur et de la fréquence des conséquences défavorables de l'IA sur les personnes et les biens. Dans l'exposé des motifs de l'*AI Act*, la Commission européenne, tout en reconnaissant les nombreuses opportunités offertes par l'IA, met en garde contre le fait que « les éléments et techniques qui rendent possibles les bénéfices socio-économiques de l'IA peuvent aussi être à l'origine de nouveaux risques ou de conséquences négatives pour les personnes ou la société ». L'*AI Act* fait suite à de nombreux travaux préliminaires à l'initiative de différents organes nationaux et européens, qui ont souligné la nécessité d'une approche équilibrée qui encouragerait le développement d'un écosystème économique et industriel favorable à l'IA sans ignorer les conséquences sur les européens. Dans cette optique, la mouture du 21 avril 2021 prétend, dans son exposé des motifs, répondre « à des demandes explicites du Parlement européen et du Conseil européen, qui ont lancé plusieurs appels en faveur de l'adoption de mesures législatives visant à assurer le bon fonctionnement du marché intérieur des systèmes d'intelligence artificielle (ci-après les « systèmes d'IA ») en mettant en balance les bénéfices et les risques

26. Certaines professions en pointe sur les conséquences de l'IA sur leur pratique du quotidien, comme l'audit et la médecine [29], mènent déjà un tel travail de prospective : notre lecture s'inscrit dans ce souci d'anticipation des risques liés à l'IA en l'élargissant aux organisations en général, en s'appuyant sur les travaux en théorie des organisations questionnant la gestion des risques et les effets de la transparence.

de l'IA à l'échelle de l'Union ».

Ni une jungle anarchique, ni un écosystème innovant étouffé par une législation inhibante : le législateur est en quête d'une troisième voix, et d'outils concrets pour la mettre en œuvre. Dans cette quête d'équilibre réglementaire, la transparence est apparue comme le principal levier normatif applicable aux organisations développant et commercialisant des IA pour matérialiser cette obligation de rendre des comptes aux nombreuses parties prenantes affectées par leurs technologies. Au fil de ses 85 articles ainsi que de ses travaux préparatoires et préambules, l'*AI Act* établit ainsi un lien clair entre exigences de transparence et gestion des risques. Les obligations de transparence sont alors modulées en fonction du niveau de risque associé au type d'IA et à son usage. Schématiquement, l'argument central derrière cette gradation des exigences en fonction des types d'IA et de leur niveau de risque repose sur l'idée que les exigences de transparence agiraient comme des garde-fous pour les organisations qui développent et commercialisent des IA : en rendant les IA plus transparentes, elles deviendraient plus explicables, ce qui faciliteraient le contrôle des acteurs. Implicitement, cette démarche repose sur l'espoir de produire trois dynamiques vertueuses. Premièrement, le contrôle rendu possible par les exigences de transparence devrait déjà produire une partie de ses fruits en obligeant les organisations et leurs acteurs à améliorer leurs processus de conception et de mise en usage en amont pour être irréprochables lorsque surviendrait l'audit. On pourrait appeler ce processus une amélioration par anticipation de la sanction. Deuxièmement, l'idéal de transparence suggère aussi que la transparence inviterait de manière corollaire l'organisation à adopter le point de vue de celles et ceux qui pourraient exiger un contrôle, et donc à développer son attention sur certains effets indésirables générés par les IA qu'elle conçoit ou qu'elle vend. On pourra ici parler d'amélioration par réflexivité organisationnelle. Enfin, en tout état de cause, le contrôle facilité par les obligations de transparence accélérerait la mise en place de plans de remédiation dans le cas où certains problèmes seraient observés : il s'agirait d'une amélioration par gain de réactivité des actions correctrices.

### Côté face : museler le doute, le tacite et l'informel, si cruciaux en matière de gestion des risques

Les efforts consentis par les organisations pour gagner en transparence se justifient donc par le gain apporté par une telle gouvernance au regard de la capacité de ces organisations à identifier et mitiger les risques en amont des contrôles (amélioration par anticipation, par réflexivité organisationnelle) et en aval (amélioration par gain de réactivité des actions correctrices). Mais dans la pratique, la littérature en gestion et de nombreuses études empiriques montrent que la transparence ne se traduit pas nécessairement par une meilleure gestion des risques. Parmi les principaux arguments invitant à la méfiance envers une application trop dogmatique de la transparence, on peut ainsi noter des effets d'occultation volontaire via la transparence ; l'effet censure de la transparence sur les savoirs tacites et informels ; et en corollaire de cette censure, l'effet rédhibitoire de la transparence sur l'expression du doute.

L'effet d'occultation par la transparence résulte d'une tentative de saturation de l'attention en inondant les observateurs d'information. Stohl *et al.* [42] distinguent entre l'opacité involontaire – dans laquelle « la visibilité produit de si grandes quantités d'informations que des éléments d'information importants sont cachés par inadvertance dans les détritus des informations rendues visibles » – et l'opacité stratégique – dans laquelle les acteurs « liés par des règles de transparence » rendent volontaire-

ment tellement d'informations « visibles que les éléments d'information sans importance prendront tellement de temps et d'efforts à passer au crible que les récepteurs seront distraits de l'information centrale que l'acteur souhaite dissimuler » (pp. 133-134).

L'effet de censure vient d'une forme d'opprobre des savoirs tacites et informels véhiculé par l'idéal de transparence dont les conséquences peuvent être plus insidieuses, et donc potentiellement plus difficiles à remédier.

L'idéal de transparence véhicule en effet l'idée que le management des risques est une affaire entièrement explicitable et objectivable. Un tel arc argumentatif se résume au fond à la sentence de Boileau : « Ce que l'on conçoit bien s'énonce clairement, Et les mots pour le dire arrivent aisément. » Dans cette logique, qualité, sécurité, éthique, répondent d'une même problématique : la nécessité d'exprimer, qui résulte forcément en une capacité à objectiver formellement le problème et ses solutions, et inversement à se méfier de toutes autres formes de gestion se fondant sur du ressenti ou de l'implicite. Or, la littérature en sciences de gestion depuis trois décennies au moins est particulièrement claire sur la nécessité de prendre en compte l'informel et le tacite comme des facteurs clé de la résilience des organisations [14, 21, 28, 41, 47]. De nombreux auteurs spécialisés sur le management des HRO (*High Reliability Organizations*[27]) ont ainsi montré comment des outils de gestion supposés identifier et objectiver les risques provoquent parfois une myopie organisationnelle sur les vraies priorités, en muselant certaines expertises et connaissances des risques fondés sur le ressenti et les savoirs informels [27, 52, 53, 54].

De peur de ne pouvoir expliquer clairement leurs raisonnements, techniciens et managers pourraient avoir tendance à taire ce qui est de l'ordre du sensible (par exemple : « j'ai la drôle d'impression que ce résultat est faux, même si je ne sais pas dire clairement pourquoi ») L'implicite hyper-rationaliste de l'idéal de transparence peut alors être contre-productif en matière de gestion des risques : sous son emprise s'opère un renversement des valeurs dans lequel le tacite est perçu comme la source du risque alors qu'il s'avère parfois pourtant crucial dans la gestion des risques : que ce soit pour permettre une identification précoces des anomalies et vulnérabilités ou pour y remédier [11, 47, 54]. Ainsi, sous le sceau de l'injonction à la transparence, les ajustements locaux, officieux, informels, sont considérés comme déviants et inacceptables : or, ils sont non seulement monnaie courante, mais même parfois franchement nécessaires à la bonne résilience des organisations [21, 47, 54]. Le bricolage et l'improvisation organisationnelle, même inexpliqués, contribuent à la bonne marche de l'organisation et au développement d'une certaine expertise [12]. En complément des plans « prescrits » de gestion des risques, la littérature souligne l'importance de respecter ce qui pourrait apparaître comme des arrangements locaux, des routines et rites, des mythes professionnels, voire des « bidouilles », du moment qu'ils participent effectivement d'une meilleure gestion de l'activité « réelle » [14, 21], sans pour autant qu'elle passe par une explicitation *in extenso* de ses tenants et aboutissants.

En mésestimant les savoirs tacites de la gestion des risques, l'injonction à la transparence pourrait alors détricoter une certaine expertise informelle en la matière, et aggraver la capacité des organisations à y faire face de manière adéquate. Mais il y a pire : corollairement au discrédit sur les savoirs tacites et

27. Les Organisations Hautement Fiables sont des organisations qui doivent parvenir à fonctionner dans des univers particulièrement complexes et incertains tout en réussissant à maintenir un haut niveau de fiabilité, comme les centrales nucléaires, les porte-avions, le contrôle aérien ou encore les urgences hospitalières.

informels, l'idéal de transparence peut tendre à assécher la pratique du doute. Or, être capable d'exprimer son doute, par exemple une crainte pour la sécurité du système fondée sur l'expérience et le ressenti, est une dimension critique de toute organisation résiliente [54]. Organiser cette reconnaissance et cette valorisation du doute nécessite non seulement d'entraîner les acteurs à identifier la sensation du doute et à le formuler à partir de leurs expériences et savoirs bien souvent tacites, mais aussi de disposer d'une culture organisationnelle et de circuits décisionnels aptes à traiter et à réagir au doute. C'est un enjeu du management de l'attention par définition bousculé, sinon écrasé, par l'idéal de transparence lorsqu'il est nourri par un présupposé hyper-rationaliste de la gestion des risques.

## Pourquoi ces problématiques organisationnelles sont importantes en matière de management de l'IA ?

En matière d'intelligence artificielle, l'expertise informelle et l'ajustement progressif sont des dimensions clés d'amélioration et de robustesse de la conception [44].

Dans cette perspective, il convient de souligner l'importance de la dimension processuelle du développement d'IA : même lorsqu'elle suit des cycles de conception rapides, l'IA n'apparait pas *ab nihilo*, et le résultat final est le cumul d'un certain nombre de décisions et de modifications dont il est impératif de garder la mémoire. En particulier, la phase d'apprentissage des algorithmes de machine learning est le fruit de successions d'ajustements et d'améliorations, guidés par un ensemble de métriques, mais aussi par de l'astuce, de l'expérience, et des jeux d'essais-erreurs qui oscillent entre ce que l'on appelle couramment les « ficelles du métier » [43] et du « bricolage » raisonné [9, 15, 37]. Les développeurs jouent sur les paramètres, les algorithmes, les bases d'apprentissage, avec une pratique relevant aussi bien des sciences computationnelles que d'un certain art.

À ce titre, ce qui compte n'est pas seulement de retracer les ajustements que les *datascientists* ont pu réaliser, mais aussi de comprendre les ajustements qu'ils n'ont pas réalisés, et de comprendre pourquoi. Glaser *et al.* [19] décrivent ainsi l'IA comme un « assemblage », combinant savoirs explicites et implicites, techniques, mais aussi relations informelles, habitudes et routines, préférences organisationnelles, modalités du processus décisionnel, etc. Une fois conçue, la phase d'adoption oblige de nouveau à modeler l'IA en fonction d'un ensemble de processus formels et informels, techniques et sociaux, propres à l'organisation adoptante. Murray *et al.* [33] décrivent ainsi une « agence siamoise » (« *conjoined agency* ») entre homme et machine, et Kellogg *et al.* [24] et Neirotti *et al.* [35] mettent en lumière les nouveaux métiers liés à l'émergence des algorithmes dans les organisations dont les missions exigent précisément d'ajouter du liant et un certain sens de l'interprétation fonctionnelle pour que les outils algorithmiques s'insèrent plus facilement dans le tissu organisationnel. Cette lecture sociotechnique et processuelle de l'IA, et en particulier le rôle de l'expérience et des ajustements durant le processus de conception, sont des dimensions qui se retrouvent non seulement occultées par l'injonction à la transparence, mais qui pourraient même à terme subir une forme de censure organisationnelle si elles ne coïncident pas avec les modalités du contrôle envisagé.

Dans cette perspective organisationnelle, deux dimensions manquantes au principe de transparence semblent ainsi nécessaires d'être renforcées dans la loi. La première porte donc sur la reconnaissance de l'informel et du tacite, incluant l'expérience et l'expertise, et l'importance du doute et des essais-erreurs dans

la gestion des risques. Si la métaphore de la « transparence », qui met l'accent sur l'objet technique, et non sur les qualifications et l'intelligence des hommes et des organisations, est insuffisante, d'autres concepts pourraient lui être favorablement substitués, comme celui de sincérité et d'auditabilité. La seconde dimension porte sur l'effet cristallisateur du principe de transparence, qui a tendance à « fixer » les limites spatiotemporelles du contrôle [6] : l'organisation est obligée de produire des rapports de l'état du système à un instant $t$. En orientant l'attention de l'organisation sur le timing du contrôle, la réglementation peut passer à côté d'un renforcement des exigences sur le processus de conception, avec ses allers-retours, ses choix, ses hésitations, ses hypothèses, etc. La mémoire de ce processus est pourtant incontournable pour une bonne gestion des risques, mais elle devient relativement incompatible avec l'eschatologie imposée en substrat par le principe de transparence. *In fine*, c'est alors tout un ressort de la résilience des organisations et du professionnalisme de ses acteurs qui pourraient s'en trouver durablement affectés.

## Références

[1] Proposition de règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (dit « AI Act »), COM/2021/206 final, 21 avril 2021. Voir la version actuelle de l'*AI Act* : Orientation générale du Conseil de l'Union européenne modifiant la proposition d'*AI Act*, document ST 15698 2022 INIT, 6 décembre 2022.

[2] M. Al Balushi. How internal transparency impacts organizational resilience. *International Journal of Quality & Reliability Management*, 38(5) :1246–1263, 2021.

[3] O. B. Albu and M. Flyverbom. Organizational transparency : Conceptualizations, conditions, and consequences. *Business & Society*, 58 :268–297, 2019.

[4] O. B. Albu and L. Ringel. The perils of organizational transparency : Consistency, surveillance, and authority negotiations. In *Toward Permeable Boundaries of Organizations ?*, pages 227–256. Emerald Publishing Limited, 2018.

[5] A. M. Almars. Deepfakes detection techniques using deep learning : a survey. *Journal of Computer and Communications*, 9(5) :20–35, 2019.

[6] M. Ananny and K. Crawford. Seeing without knowing : Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3) :973–989, 2018.

[7] G. Andrada, R. W. Clowes, and P. R. Smart. Varieties of transparency : Exploring agency within ai systems. *AI & Society*, pages 11–11, 2022.

[8] E. Bietti. From ethics washing to ethics bashing : a view on tech ethics from within moral philosophy. In *Conference on fairness, accountability, and transparency*, pages 210–219, 2020.

[9] E. Boxenbaum and L. Rouleau. New knowledge products as bricolage : Metaphors and scripts in organizational theory. *Academy of Management Review*, 36(2) :272–296, 2011.

[10] A. Buhmann and C. Fieseler. Towards a deliberative framework for responsible innovation in artificial intelligence. *Technology in Society*, 64 :101475, 2011.

[11] D. Chernov, D. Sornette, G. Sansavini, and A. Ayoub. Examples of failures in intra-organizational risk transmission in past disasters. In *Don't Tell the Boss !*

*How Poor Communication on Risks within Organizations Causes Major Catastrophes*, pages 13–332. Springer International Publishing, 2023.

[12] F. Chédotel. L'improvisation organisationnelle. *Revue française de gestion*, 1 :123–140, 2005.

[13] K. de Fine Licht and J. de Fine Licht. Artificial intelligence, transparency, and public decision-making : Why explanations are key when trying to produce perceived legitimacy. *AI & Society*, 35 :917–926, 2020.

[14] M. Detchessahar, S. Gentil, A. Grevin, and B. Journé. Entre cacophonie et silence organisationnel, concevoir le dialogue sur le travail. le cas de projets de maintenance dans une industrie à risque. *Annales des Mines – Gérer et comprendre*, 4 :33–45, 2017.

[15] R. Duymedjian and C. C. Rüling. Towards a foundation of bricolage in organization and management theory. *Organization Studies*, 31(2) :133–151, 2017.

[16] L. Edwards and M. Veale. Slave to the algorithm ? why a 'right to an explanation' is probably not the remedy you are looking for. *16 Duke Law & Technology Review*, 18 :2972855, 2017.

[17] N. Emaminejad and R. Akhavian. Trustworthy AI and robotics : Implications for the AEC industry. *Automation in Construction*, 139 :104298, 2022.

[18] H. Felzmann, E. F. Villaronga, C. Lutz, and A. Tamò-Larrieux. Transparency you can trust : Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, 6(1) :2053951719860542, 2019.

[19] V. L. Glaser, N. Pollock, and L. D'Adderio. The biography of an algorithm : Performing algorithmic technologies in organizations. *Organization Theory*, 2(2) :26317877211004609, 2021.

[20] T. Hagendorff. The ethics of AI ethics : An evaluation of guidelines. *Minds and Machines*, 30(1) :99–120, 2020.

[21] E. Hollnagel, B. Journé, and H. Laroche. La fiabilité et la résilience comme dimensions de la performance organisationnelle. *M@n@gement*, 12(4) :224–229, 2009.

[22] M. Jendly. Performance, transparence et accountability : une équation (dé) responsabilisante des professionnels exerçant en prison ? *Déviance et société*, 36(3) :243–262, 2012.

[23] A. Jobin, M. Ienca, and E. Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9) :389–399, 2019.

[24] K. C. Kellogg, M. A. Valentine, and A. Christin. Algorithms at work : The new contested terrain of control. *Academy of Management Annals*, 14(1) :366–410, 2020.

[25] S. Larsson and F. Heintz. Transparency in artificial intelligence. *Internet Policy Review*, 9(2) :2020.2.1469, 2020.

[26] K. Martin. Ethical implications and accountability of algorithms. *Journal of Business Ethics*, 160 :835–850, 2019.

[27] J. Mayer. *De l'attention au risque : une perspective attentionnelle de la construction sociale du risque par les organisations*. PhD thesis, Université Paris-Dauphine – Paris Sciences et Lettres, 2017.

[28] J. C. Mayer. Influencer l'attention des décideurs-les pratiques d'« issue-selling » des risk managers. *Revue française de gestion*, 42(255) :75–88, 2016.

[29] S. McLennan, A. Fiske, D. Tigard, R. Müller, S. Haddadin, and A. Buyx. Embedded ethics : a proposal for integrating

ethics into the development of medical AI. *BMC Medical Ethics*, 23(1) :6, 2022.

[30] B. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi. The ethics of algorithms : Mapping the debate. *Big Data & Society*, 3(2) :2053951716679679, 2016.

[31] M. Mora-Cantallops, S. Sánchez-Alonso, E. García-Barriocanal, and M. A. Sicilia. Traceability for trustworthy ai : A review of models and tools. *Big Data and Cognitive Computing*, 5(2) :20, 2021.

[32] J. Morley, A. Elhalal, F. Garcia, L. Kinsey, J. Mökander, and L. Floridi. Ethics as a service : a pragmatic operationalisation of AI ethics. *Minds and Machines*, 31(2) :239–256, 2021.

[33] A. Murray, J. E. N. Rhymer, and D. G. Sirmon. Humans and technology : Forms of conjoined agency in organizations. *Academy of Management Review*, 46(3) :552–571, 2021.

[34] J. Mökander and L. Floridi. Operationalising AI governance through ethics-based auditing : an industry case study. *AI and Ethics*, s43681-022-00171-7, 2022.

[35] P. Neirotti, D. Pesce, and D. Battaglia. Algorithms for operational decision-making : An absorptive capacity perspective on the process of converting data into relevant knowledge. *Technological Forecasting and Social Change*, 173 :121088, 2021.

[36] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M. E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernandez, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis, and S. Staab. Bias in data-driven artificial intelligence systems – an introductory survey. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, 10(3) :e1356, 2020.

[37] N. Phillips and P. Tracey. Opportunity recognition, entrepreneurial capabilities and bricolage : connecting institutional theory and entrepreneurship in strategic organization. *Strategic Organization*, 5(3) :313–320, 2007.

[38] L. Ringel. Boundaries of visibility in the age of transparency : An integrative conceptualization. In *Toward Permeable Boundaries of Organizations ?*, pages 55–79. Emerald Publishing Limited, 2018.

[39] L. Ringel. Unpacking the transparency-secrecy nexus : Frontstage and backstage behaviour in a political party. *Organization Studies*, 40(5) :705–723, 2019.

[40] J. Spindler. La transparence de la gestion publique : de la recherche d'un plus grand approfondissement à un risque d'opacité. *Gestion Finances Publiques*, 6(6) :68–77, 2020.

[41] A. Stimec and B. Journé. Faire face à la complexité par le dialogue technique et la négociation-le cas de la sûreté dans l'industrie nucléaire. *Revue française de gestion*, 47(297) :123–140, 2021.

[42] C. Stohl, M. Stohl, and P. M. Leonardi. Managing opacity : Information visibility and the parado of transparency in the digital age. *International Journal of Communication*, 10 :123–137, 2016.

[43] M. Stroobants. Dénouer les ficelles du métier pour connecter les savoirs formels et informels. *Techniques & Culture*, 51 :164–179, 2009.

[44] L. A. Suchman and R. H. Trigg. Artificial intelligence as craftwork. In S. Chaiklin and J. Lave, editors, *Understanding practice : Perspectives on activity and*

*context*, pages 144–178. Cambridge University Press, 1993.

[45] B. Van der Sloot and Y. Wagensveld. Deepfakes : regulatory challenges for the synthetic society. *Computer Law & Security Review*, 46 :105716, 2022.

[46] M. Veale and F. Zuiderveen Borgesius. Demystifying the draft EU Artificial Intelligence act (july 31, 2021). *Computer Law Review International*, 22(4) :97–112, 2021.

[47] T. J. Vogus, N. B. Rothman, K. M. Sutcliffe, and K. E. Weick. The affective foundations of high-reliability organizing. *Journal of Organizational Behavior*, 35(4) :592–596, 2014.

[48] E. S. Vorm and D. J. Combs. Integrating transparency, trust, and acceptance : The intelligent systems technology acceptance model. *International Journal of Human-Computer Interaction*, 38(18–20) :1828–1845, 2022.

[49] J. Wanner, L. V. Herm, K. Heinrich, and C. Janiesch. The effect of transparency and trust on intelligent system acceptance : Evidence from a user-based study. *Electronic Markets*, 32 :2079–2102, 2022.

[50] R. Warner and R. H. Sloan. Making artificial intelligence transparent : Fairness and the problem of proxy variables. *Criminal Justice Ethics*, 40(1) :23–39, 2021.

[51] A. L. Washington. How to argue with an algorithm : Lessons from the COMPAS-ProPublica debate. *Colorado Technology Law Journal*, 17(1) :3357874, 2019.

[52] K. E. Weick and K. M. Sutcliffe. *Managing the unexpected*. San Francisco : Jossey-Bass, 2001.

[53] K. E. Weick and K. M. Sutcliffe. Mindfulness and the quality of organizational attention. *Organization Science*, 17(4) :514–524, 2006.

[54] K. E. Weick and K. M. Sutcliffe. *Managing the unexpected : Resilient performance in an age of uncertainty*. John Wiley & Sons, 2007.

[55] J. Zerilli, U. Bhatt, and A. Weller. How transparency modulates trust in artificial intelligence. *Patterns*, 3 :100455, 2022.

AfIA
Association française
pour l'Intelligence Artificielle

# Quel rôle pour les mathématiques dans le traitement des problèmes d'équité en IA ?

Par

**Evgenii CHZHEN**
*LMO*
*CNRS, Université Paris-Saclay*
*evgenii.chzhen@cnrs.fr*

**Christophe DENIS**
*LAMA*
*Université Gustave Eiffel*
*christophe.denis@univ-eiffel.fr*

**Mohamed HEBIRI**
*LAMA*
*Université Gustave Eiffel*
*mohamed.hebiri@univ-eiffel.fr*

**Nicolas SCHREUDER**
*DIBRIS, MaLGa*
*Università di Genova*
*schreuder.nicolas@gmail.com*

**Gayane TATURYAN**
*IRT SystemX*
*Université Gustave Eiffel*
*gayane.taturyan@irt-systemx.fr*

## Introduction

Cette contribution présente la problématique de l'équité des algorithmes prédictifs d'IA du point de vue de la statistique mathématique. Nous commençons par discuter de la nécessité d'une formalisation mathématique du problème, avant de présenter les principales techniques pour contrôler les biais indésirables de prédiction. Finalement, nous décrivons quelques résultats statistiques représentatifs pour quantifier et contrôler les biais d'une de ces approches.

## Problématique

Notre vie quotidienne est de plus en plus influencée par des prédictions algorithmiques. Il est désormais admis que ces prédictions peuvent reproduire, voire amplifier, certaines discriminations de notre société, notamment en raison du processus d'apprentissage des algorithmes d'IA [2]. Par exemple, les ingénieurs d'Amazon ont découvert que leur algorithme de présélection de candidats était négativement biaisé envers les femmes, sans que ce biais ait été introduit intentionnellement. L'algorithme, entraîné sur des données historiques, perpétuait un déséquilibre existant entre les hommes et les femmes dans le secteur des nouvelles

28. « Amazon scraps secret AI recruiting tool that showed bias against women » (Reuters, en anglais).

technologies. Admettant un manque de compréhension de ce problème, les équipes de recherche d'Amazon n'ont pas été en mesure de le résoudre et ont décidé de mettre de côté cet algorithme [28]. Plus généralement, l'automatisation croissante des procédures de prise de décision – des prêts bancaires aux admissions scolaires, en passant par les condamnations pénales – augmente considérablement le risque d'automatiser simultanément les discriminations. En France, cette menace a récemment conduit la CNIL et le Défenseur des droits à recommander de « soutenir la recherche pour développer les études de mesure et de prévention des biais, et approfondir la notion de *fair learning* [29]. »

L'*équité algorithmique* (ou *fair learning*) est un domaine de recherche se situant à l'intersection de différentes disciplines (entre autres, les mathématiques, les statistiques, le droit, la sociologie et l'informatique), dont l'objectif est de comprendre les phénomènes pouvant potentiellement rendre les prédictions algorithmiques *inéquitables* et d'en contrôler l'impact.

Dans cette contribution, nous nous focaliserons sur les aspects mathématiques et statistiques de cette problématique. Dans un premier temps, nous discuterons de la nécessité de formaliser mathématiquement ce problème. Dans un second temps, nous présenterons les principales stratégies techniques existantes pour corriger les biais et contrôler les discriminations résultants des prédictions algorithmiques. Finalement, nous donnerons un aperçu de résultats que nous avons obtenu pour quantifier et certifier la performance et l'équité d'une procédure de prédiction dans un cadre donné.

## Formalisation mathématique de l'équité

Un algorithme d'apprentissage est, par définition, discriminant : les prédictions algorithmiques sont basées sur l'identification de motifs statistiques – des corrélations – qui relient un ensemble de caractéristiques observables à un phénomène à prédire. Il existe deux phases principales dans la conception et l'utilisation d'un algorithme de prédiction pour une tâche donnée. Au cours de la première, la phase d'apprentissage, un processus d'optimisation mathématique est implémenté pour identifier, à partir d'une base de données historiques, les corrélations entre un ensemble de variables dites prédictives et une variable à prédire. L'objectif de cette phase est de distinguer les corrélations utiles pour la tâche de prédiction des corrélations inutiles et/ou fallacieuses. Lors de la seconde phase, dite de prédiction, l'algorithme infère la valeur de la variable à prédire d'une nouvelle observation à partir des valeurs des variables prédictives.

Pour clarifier cette description, prenons l'exemple de la prédiction des salaires dans une organisation. Nous disposons d'une base de données de profils et de salaires d'employés sur le marché du travail d'un secteur donné et nous aimerions prédire des niveaux de salaire « justes » et personnalisés pour de nouvelles recrues. Pour ce faire, nous pouvons entraîner un algorithme d'apprentissage sur cette base de données. Si, lors de la phase d'apprentissage, le processus d'optimisation met en lumière une corrélation entre le nombre d'années d'expérience professionnelle et le salaire, une nouvelle recrue bénéficiant de nombreuses années d'expérience se verra attribuer un salaire plus élevé qu'une autre recrue moins expérimentée *ceteris paribus*. De la même manière, s'il existe un écart de salaire entre les hommes et les femmes dans la base de données, il est possible que celui-ci soit identifié puis reproduit par l'algorithme lors de la phase de prédiction.

Nous voyons ici que certaines corrélations sont utiles et justifiables – par exemple, celle

29. Algorithmes : prévenir l'automatisation des discriminations.

entre le nombre d'années d'expériences et le salaire – tandis que d'autres, même si elles permettent d'améliorer la capacité prédictive d'un algorithme par rapport à un jeu de données, peuvent être indésirables.

Du point de vue éthique, la question fondamentale est de déterminer quelles corrélations peuvent être exploitées par des algorithmes de prédiction et quels biais doivent être corrigés. Les réponses à cette question doivent nécessairement être traduites dans le langage mathématique pour pouvoir les communiquer aux algorithmes – des entités mathématiques conçues à partir de la théorie mathématique.

Il existe différentes manières de formaliser mathématiquement l'équité d'une règle de prédiction [2], chacune correspondant à un choix de valeurs pour un contexte donné. Nous nous focaliserons ici sur la notion de *Parité Démographique*. Formellement, une règle de décision satisfait le critère de Parité Démographique si et seulement si la distribution de la prédiction est statistiquement indépendante des attributs sensibles contre lesquels nous souhaitons contrôler d'éventuelles discriminations. Dans l'exemple des salaires, une règle de prédiction de salaire satisfait cette contrainte si la distribution des salaires prédits pour les hommes et pour les femmes coïncident à l'échelle de la population. Autrement dit, la connaissance du salaire prédit ne révèle aucune information sur le genre de l'individu considéré.

Bien que l'équité de l'algorithme soit un objectif en soi, il est crucial d'élaborer des règles de prédiction en accord avec les objectifs habituels de performance. En effet, dans l'exemple ci-dessus, une prédiction trop élevée par rapport au marché coûterait trop cher à l'entreprise et une prédiction trop basse lui ferait perdre de nombreux candidats potentiels.

Une fois le critère d'équité formalisé mathématiquement, se pose donc le problème de la conception d'algorithmes qui satisfont ce critère et qui préservent une bonne capacité prédictive. La section suivante présente différentes stratégies pour contrôler le caractère potentiellement discriminatoire et inéquitable des règles de prédictions.

## Stratégies de réduction de biais

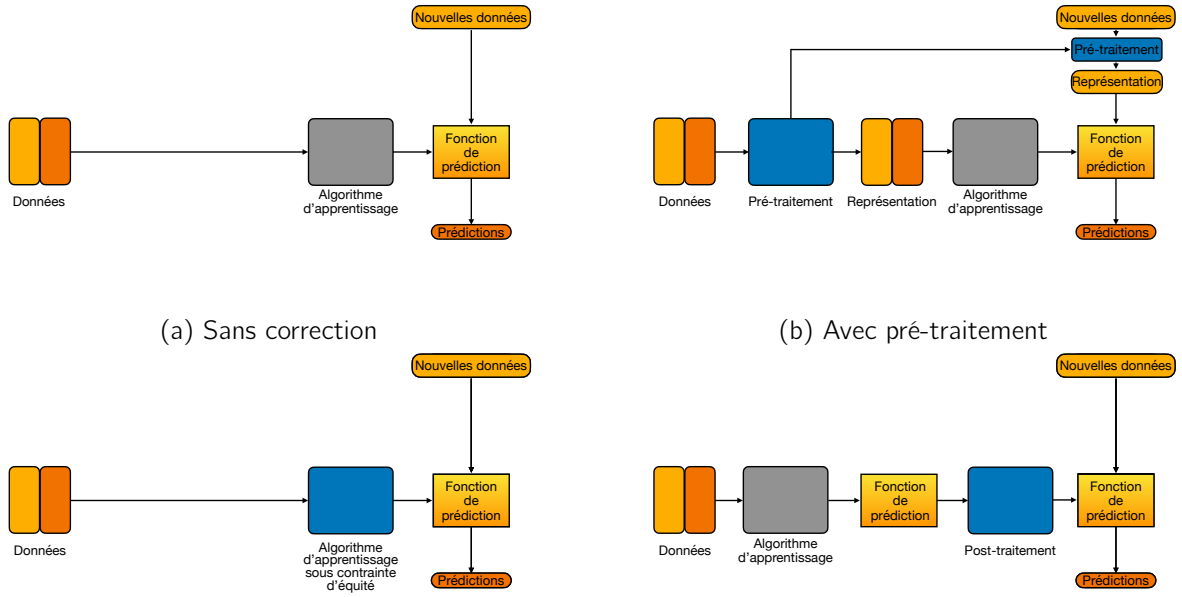
Une idée naturelle, mais erronée, pour éliminer de potentielles discriminations d'une règle de prédiction est de ne pas révéler certaines caractéristiques sensibles (par exemple, le genre ou l'ethnie) à l'algorithme d'apprentissage en les effaçant des données.

L'existence de corrélations entre des caractéristiques sensibles cachées et des caractéristiques accessibles à l'algorithme induit une corrélation entre la prédiction de l'algorithme et ces caractéristiques sensibles inaccessibles. Ainsi, alors même que l'algorithme n'a pas accès aux variables sensibles, la règle de décision qui en résulte peut être biaisée vis-à-vis de ces variables, ce qui peut constituer une discrimination indirecte[30]. Il est donc nécessaire de développer d'autres approches pour concevoir des d'algorithmes équitables. Nous détaillons ci-après les trois principales approches connues.

Une première approche, dite de *pre-processing*, consiste à dé-biaiser les données en corrigeant directement leurs éventuels biais [9, 10, 8]. Cette approche consiste en l'apprentissage d'une représentation des données pour laquelle le biais est atténué (cf. figure 1-(b)). Un algorithme d'IA peut ensuite être entraîné de manière usuelle sur cette nouvelle représentation des données.

Il existe deux autres approches pour réduire les biais des règles de décisions, celles-ci concernent directement les règles de décision et non pas les données. La première approche,

30. Voir en particulier l'article 1 de la loi n° 2008-496 du 27 mai 2008 portant diverses dispositions d'adaptation au droit communautaire dans le domaine de la lutte contre les discriminations.




(a) Sans correction

(b) Avec pré-traitement

(c) Apprentissage sous contrainte

(d) Avec post-traitement

Figure 1 – Représentations schématiques

dite d'*in-processing* (cf. figure 1-(c)), consiste à intégrer le critère d'équité comme contrainte explicite dans la phase d'apprentissage [1, 6]. La seconde approche, dite de *post-processing* (cf. figure 1-(d)), consiste à modifier les prédictions d'un algorithme existant pour en corriger les biais [7, 3].

En pratique, l'approche *in-processing* peut s'avérer coûteuse car elle nécessite de changer intégralement l'architecture d'un système algorithmique. Les deux autres approches ont pour avantage de s'insérer dans des systèmes pré-existants, facilitant leur implémentation. De plus, contrairement à l'approche *pre-processing*, l'approche *post-processing* ne nécessite pas de réapprendre les paramètres de l'algorithme, le format de données en entrée restant inchangé.

## Apports de la théorie statistique

Dans la section précédente, nous avons introduit les principales approches étudiées et implémentées pour contrôler les biais et discriminations des algorithmes de prédiction. Chacune consiste en un ensemble de modifications des processus usuels d'apprentissage et de prédiction, donnant lieu à de nouveaux algorithmes. Pour que ces modifications ne soient pas vaines – ou même nocives – il est crucial d'évaluer la capacité de ces nouveaux algorithmes à remplir la tâche pour lesquelles ils ont été conçus. Nous pourrions étudier empiriquement le comportement d'algorithmes sur des jeux de données réels, mais rien ne garantirait la généralité des phénomènes ainsi observés. Dès lors, comment *certifier* que les approches et algorithmes proposés remplissent leurs objectifs d'équité et de performance ? Pour ce faire, il est nécessaire de développer une théorie de l'équité des algorithmes qui prenne en compte la variabilité des

tâches considérées et des données observées.

Étant donné des mesures de performance prédictive et d'équité, notre objectif en tant que statisticiens est (i) de quantifier les meilleurs compromis, entre capacité prédictive et équité, atteignables par les algorithmes de prédiction (ii) de concevoir des algorithmes qui réalisent ces compromis. Il est important de souligner que notre approche n'aborde pas la pertinence d'un choix donné (comme le choix d'un critère d'équité ou d'une mesure de performance), qui est finalement laissée au décideur, mais permet de mieux comprendre les conséquences de ce choix. Nous présentons maintenant deux résultats représentatifs de notre travail sur l'équité des algorithmes du point de vue de la théorie statistique.

Tout d'abord, une analyse mathématique du problème de prédiction sous contrainte de non-discrimination fournit une description explicite de la fonction de prédiction optimale – au sens de la capacité prédictive – parmi celles qui satisfont la contrainte de Parité [4]. Cette description permet de comprendre précisément quels types de discrimination peuvent être contrôlés par la règle de décision équitable (au sens de la Parité Démographique) et optimale.

La règle de décision obtenue est idéale dans le sens où elle dépend de quantités théoriques inaccessibles en pratique. Néanmoins, elle a servi de point de départ à la construction d'estimateurs statistiques – c'est-à-dire de fonctions calculables à partir de données observées – pour lesquels nous avons obtenu des garanties sur la capacité prédictive et la non-discrimination.

Ensuite, [5] a étendu ces résultats à une notion de Parité Démographique approchée. Ce travail fournit une caractérisation des solutions Pareto-optimale pour le problème du contrôle simultané des capacités prédictives et de non-discriminations. Cette caractérisation permet de répondre précisément à des questions telles que : quel est le niveau minimal de discrimination atteignable pour un seuil d'erreur de prédiction donné ?

## Conclusion

Le contrôle des biais discriminatoires issus de prédictions algorithmiques est un enjeu crucial pour notre société.

Pour atteindre cet objectif, il est nécessaire de disposer d'outils pour identifier et quantifier ces biais. La théorie de l'apprentissage statistique permet de développer de tels outils et de concevoir de règles de décision équitables selon un critère formel.

Dans cette contribution, nous nous sommes focalisés sur une notion mathématique d'équité, la Parité Démographique, parmi un ensemble de notions étudiées dans la littérature. Nous avons montré que l'analyse mathématique permet de fournir une description objective des conséquences du choix de ce critère. En particulier, nous n'avons pas discuté du bien-fondé de ce choix d'un point de vue éthique. Nous insistons sur le fait qu'un dialogue interdisciplinaire est nécessaire avant tout déploiement d'algorithmes qui satisfont des contraintes d'équité. Nous espérons que notre analyse permettra de clarifier les débats interdisciplinaires qui émergent autour de la problématique de l'équité des algorithmes.

## Références

[1] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *37th International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

[2] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning : Limitations and Opportunities*. fairmlbook.org, 2019.

[3] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. *Advances in Neural Information Processing Systems*, 32, 2019.

[4] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Fair regression with wasserstein barycenters. *Advances in Neural Information Processing Systems*, 33 :7321–7331, 2020.

[5] E. Chzhen and N. Schreuder. A minimax framework for quantifying risk-fairness trade-off in regression. *The Annals of Statistics*, 50(4) :2416–2442, 2022.

[6] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. *Advances in Neural Information Processing Systems*, 31, 2018.

[7] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

[8] L. Oneto, M. Donini, G. Luise, C. Ciliberto, A. Maurer, and M. Pontil. Exploiting mmd and sinkhorn divergences for fair and transferable representation learning. *Advances in Neural Information Processing Systems*, 33 :15360–15370, 2020.

[9] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *30th International Conference on Machine Learning*, pages 325–333, 2013.

[10] H. Zhao and G. Gordon. Inherent tradeoffs in learning fair representations. *Advances in Neural Information Processing Systems*, 32, 2019.

AfIA
Association française
pour l'Intelligence Artificielle

# Recherche en éthique computationnelle au sein de l'équipe ACASA du Lip6 – Cadre ACE de modélisation des raisonnements éthiques

Par

**Gauvain BOURGNE**
*Lip6 / ACASA*
*Sorbonne Université*
*gauvain.bourgne@lip6.fr*

**Jean-Gabriel GANASCIA**
*Lip6 / ACASA*
*Sorbonne Université*
*jean-gabriel.ganascia@lip6.fr*

**Camilo SARMIENTO**
*Lip6 / ACASA*
*Sorbonne Université*
*camilo.sarmiento@lip6.fr*

**Yousef TAHERI**
*Lip6 / ACASA*
*Sorbonne Université*
*yousef.taheri@lip6.fr*

**Guillaume GERVOIS**
*Lip6 / ACASA*
*Sorbonne Université*
*guillaume.gervois@lip6.fr*

## Introduction

Au sein du Lip6, dans l'équipe ACASA, nous développons depuis plusieurs années une réflexion sur la modélisation des raisonnements éthiques qui nous a amenés à proposer le cadre ACE, un cadre modulaire implémenté en ASP permettant l'évaluation d'actions selon différents principes éthiques [3, 6]. Ce cadre est basé sur une décomposition en trois couches dont il tire son nom : Action-Causalité-Éthique. Il a été développé à l'occasion des travaux de thèse de Fiona BERREBY [1] au sein du projet ANR ETHICAA sur l'éthique des agents autonomes. Il se prolonge aujourd'hui notamment par plusieurs thèses en cours et la participation au projet trilatéral franco-allemand-japonais ANR-DFG-JST ReComp dédié à la vérification de conformité à la fois légale et éthique.

Cet article présente les principes et la structure de ce cadre, en commençant par le situer dans le domaine de l'éthique computationnelle.

## Éthique computationnelle

L'éthique computationnelle est un domaine de recherche pluridisciplinaire, entre informatique, philosophie et sociologie, qui s'attache à utiliser les outils de l'informatique et de d'intelligence artificielle pour étudier ou traduire les mécanismes à l'œuvre dans les problèmes éthiques. Il se distingue en cela de la cyber-éthique, champs de l'éthique appliquée, qui étudie les implications éthiques des usages de l'intelligence artificielle et de l'informatique.

C'est un domaine large qui comporte de nombreuses approches selon l'objet ou la méthode. Si l'essentiel des travaux s'intéresse à la question de la prise de décision éthique, il existe aussi des approches agissant plutôt en amont, en s'intéressant à la phase de conception d'algorithme (*ethics by design*) ou en aval, en développant des mécanismes de vérification.

Une division classique des approches de décision éthique différencie les approches prescriptives, se fondant sur les théories d'éthique normative pour formaliser des principes, et les approches descriptives, tentant de capturer le sens moral commun, souvent à l'aide de méthodes d'apprentissage. Nos travaux s'inscrivent résolument dans une approche prescriptive, en s'appuyant sur des formalismes logiques. Ce choix se fonde sur un souci d'expliciter les règles employées et de maintenir un mécanisme « boite blanche » qui nous paraît indispensable sur le sujet de l'éthique.

Les travaux initiaux de cette branche se sont souvent focalisés sur des modèles spécifiques tels que l'utilitarisme de l'acte [13], la doctrine du double effet [2, 5, 10] ou l'impératif catégorique de Kant [12]. Différents formalismes ont été employés, de la programmation logique au calcul des évènements déontiques [10], avec en conséquence différentes représentations incompatibles des dilemmes ou exemples classiques dans ces formalismes.

Partant de ce constat, nos travaux ont l'ambition de rassembler et comparer les différents principes formalisés en proposant un cadre général et une représentation unifiée des exemples. Cet objectif se retrouve aussi dans d'autres travaux récents [8, 11]. En particulier, l'approche HERA [12, 13], qui se fonde sur des modèles causaux, s'inscrit dans la même démarche que la nôtre.

Bien que nous nous limitions pour le moment au raisonnement individuel d'un agent, l'extension à une dimension collective pose de nombreux problèmes intéressants et des travaux existent déjà dans ce domaine [7].

## Le raisonnement éthique

Nous nous intéressons à la modélisation des raisonnements éthiques. Les travaux en éthique normative s'attellent à fournir des principes pour permettre des décisions justes. On distingue généralement trois grandes approches selon que l'on s'appuie sur le bien produit par les conséquences de la décision (approche conséquentialiste), sur le respect d'un devoir formalisé en règle de comportement (approche déontologique) ou sur des considérations plus internes liées au développement des vertus du décideur (éthique de la vertu).

Il s'agit donc pour notre cadre de modéliser des mécanismes de prise de décision éthique selon ces différents principes. Nous nous concentrons sur l'évaluation du *juste*, qui revient à filtrer, ordonner ou valuer différentes options qui se présentent. L'essentiel des théories d'éthique normative se concentre plutôt sur des mécanismes de filtrage en définissant les conditions selon lesquelles une action est impermissible ou permissible. Le raisonnement éthique a alors pour objet l'évaluation de la permissibilité éthique d'une décision et c'est la tâche que notre cadre se propose de formaliser.

Une décision peut prendre différentes formes selon sa portée : d'une simple action simple à une politique générale, en passant par des plans séquentiels ou contingents. L'essentiel des dilemmes et expériences de pensée mis en avant dans le champ philosophique considère des situations et des actions spécifiques où la décision est de l'ordre d'un choix entre plusieurs actions simples. Notre étude initiale se focalise donc sur l'évaluation d'une action, avec des extensions naturelles pour considérer des plans. L'évaluation de politiques générales de décision diffère dans la forme des problèmes et reste dans le champ des travaux futurs.

## Principes du cadre ACE

Pour pouvoir représenter plusieurs principes dans le même cadre afin de les mettre en rapport et de les comparer, il est nécessaire d'avoir une représentation commune des aspects factuels de la situation évaluée. En ce sens, le cadre ACE adopte un principe de séparation des modèles à différents niveaux.

Tout d'abord, il est essentiel d'identifier et de séparer les aspects éthiques des aspects plus factuels. La représentation de la situation et de son évolution en fonction des actions possibles ne dépend en rien des choix éthiques et doit donc faire l'objet d'un modèle indépendant sur lequel pourra se fonder l'évaluation éthique. Cette séparation permet aussi d'identifier les questions de représentation propres aux principes modélisés.

Le second axe de notre principe de séparation des modèles consiste à clairement séparer les connaissances modélisées selon leur généralité, différencier le général du spécifique. À la manière des modèles de planification, nous séparons les faits spécifiques à une situation ou un problème donné, les mécanismes propres au domaine modélisé et les axiomes généraux qui régissent tous les domaines. Sur les aspects éthiques, cela permet de dégager les règles générales permettant de définir les différents principes en identifiant les spécifications propres à la situation ou au domaine qui sous-tendent ces axiomes.

Enfin, le cadre proposé est un cadre modulaire et implémenté. Chaque brique du raisonnement peut faire l'objet de variantes ou d'implémentations indépendantes : l'identification des entrées et sorties de chaque module permettant de repérer les interdépendances. L'ensemble du cadre est implémenté en *Answer Set Programming* et le code des différentes versions du cadre est disponible sur le Gitlab du Lip6 ici et ici.

## Structure Action-Causalité-Éthique

Le cadre ACE propose une structure en trois niveaux : (i) une couche factuelle de description de la situation et de ses évolutions en fonction des décisions, (ii) une seconde couche plus abstraite d'analyse des relations, centrée sur la causalité, et enfin, (iii) une couche éthique qui s'appuie sur cela pour évaluer la permissibilité éthique des options considérées.

La première couche est le modèle d'action. Elle a pour objet de modéliser factuellement la situation et ses évolutions possibles. Tirant parti des travaux sur la représentation de l'action et du changement, elle s'appuie sur les clarifications de concept opérées dans le domaine de la planification et notamment illustrées par le formalisme PDDL. L'étude des dilemmes classiques a rapidement fait apparaître que s'il était possible de rester dans un cadre de planification classique (déterministe et sans évènements duratifs), il était utile de prendre en compte des évènements exogènes ou déclenchables et de considérer la possibilité d'actions concurrentes. Le premier modèle [2, 3, 6], formalisé comme fragment réduit du calcul des évènements et implémenté en ASP, s'appuie sur le fragment STRIPS de PDDL enrichi d'actions exogènes (appelés évènements automatique dans notre cadre). Ce modèle est constitué d'un contexte et d'un moteur d'action. Le contexte est composé d'un modèle du domaine (réifié en un ensemble de faits sur les propriétés des évènements) et d'une situation initiale. Les options envisagées sont données comme un ensemble de scénarios. Le modèle d'action détermine pour chacun de ces scénarios l'évolution des états et les évènements déclenchés à chaque pas de temps.

La seconde couche est le modèle causal. Établir les liens causaux est en effet nécessaire à un grand nombre de raisonnements éthiques. Les approches conséquentialistes s'appuient en

effet sur une évaluation des conséquences de l'action, et s'appuyer sur des relations causales fondées plutôt que sur une simple constatation de l'état final devient nécessaire si les acteurs sont multiples. L'établissement de relations causales est aussi indispensable pour tous les principes s'appuyant sur des considérations de moyens pour une fin, tels que la doctrine du double effet, qui stipule que lorsqu'une action produit à la fois une conséquence bonne et une mauvaise, la bonne conséquence ne doit pas découler de la mauvaise. Un autre exemple est celui de la seconde formulation de l'impératif catégorique de Kant, qui exige de traiter l'humanité toujours en même temps comme fin et jamais simplement comme moyen. Nous adoptons pour cela une approche fondée sur les évènements [4], s'appuyant sur les effets et préconditions des actions et évènements déclenchables. Ce modèle établit des relations causales aussi bien positives (causer un évènement) que négatives (empêcher un évènement).

Pour finir, la couche éthique utilise les analyses des couches précédentes pour évaluer les actions permissibles. Chaque principe éthique peut se formaliser en une théorie du Juste qui détermine la permissibilité des actions. Nous identifions pour chaque principe les spécifications éthiques nécessaires à compléter l'évaluation, c'est-à-dire les connaissances éthiques permettant d'appliquer la règle générale à la situation particulière considérée. Dans le cas d'approche conséquentialistes, cela prend la forme d'une théorie du Bien évaluant les conséquences dans le domaine modélisé. Pour les approches déontologique, il peut être nécessaire de définir les règles de conduite propre au domaine ou d'identifier les fins visées par l'action.

## Travaux en cours et perspectives

Les travaux sur ce cadre se prolongent actuellement, en particulier à travers trois thèses.

Un premier axe d'amélioration consiste en une formalisation sémantique des différentes étapes du raisonnement, couplée à un enrichissement de l'expressivité du modèle d'action. Dans son travail de thèse, Camilo SARMIENTO propose un langage d'action permettant évènements exogènes, actions concurrentes et expressivité accrue (avec préconditions disjonctives et effets conditionnels) sur la base duquel une définition sémantique de la causalité est proposée [15]. Une implémentation correcte et complète en ASP est aussi proposée [14]. Les perspectives actuelles concernent l'analyse fine des situations causales de surdétermination et la prise en compte d'absences d'évènement dans les relations causales, aussi bien comme causes (modéliser les omissions d'actions [6]) que comme conséquences (relations causales de préventions).

Dans le cadre du projet ReComp, à travers notamment la thèse de Yousef TAHERI, nous nous intéressons à l'intégration de mécanismes de conformité légale [16] et d'évaluation éthique à une architecture de planification par réseau hiérarchique de tâches (HTN – *Hierarchical Task Network*). En s'intéressant au domaine de la gestion des données personnelles, qui soulève aussi bien des problématiques légales qu'éthiques, nous sommes amenés à étudier des principes conséquentialistes pluralistes, permettant de modéliser et prendre en compte une hiérarchie de valeurs telles que celles mises en avant par les comités d'éthique sur l'IA de confiance.

Enfin, la thèse de Guillaume GERVOIS, commencée en octobre 2022 et co-encadrée par Marie-Jeanne LESOT de l'équipe LFI du Lip6, s'intéresse à une approche ordinale du raisonnement éthique, en étudiant la compatibilité des relations d'ordre entre les actions induites par différents principes. Des premiers travaux sur les questions de supériorité entre différentes modalités de bien ont déjà fait l'ob-

jet d'une première publication [9].

## Références

[1] F. Berreby. *Models of ethical reasoning*. PhD thesis, Sorbonne Université, Paris, September 2018.

[2] F. Berreby, G. Bourgne, and J.-G. Ganascia. Modelling moral reasoning and ethical responsibility with logic programming. In *20th International Conference, Logic for Programming, Artificial Intelligence, and Reasoning*, volume 9450, pages 532–548, 2015.

[3] F. Berreby, G. Bourgne, and J.-G. Ganascia. A declarative modular framework for representing and applying ethical principles. In *16th International Conference on Autonomous Agents and MultiAgent Systems*, pages 96–104, 2017.

[4] F. Berreby, G. Bourgne, and J.-G. Ganascia. Event-based and scenario-based causality for computational ethics. In *17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 147–155, 2018.

[5] V. Bonnemains, C. Saurel, and C. Tessier. Embedded ethics : some technical and ethical challenges. *Ethics and Information Technology*, 20(1) :41–58, 2018.

[6] G. Bourgne, C. Sarmiento, and J.-G. Ganascia. ACE modular framework for computational ethics : dealing with multiple actions, concurrency and omission. In *1st International Workshop on Computational Machine Ethics*. CEUR-WS.org, 2021.

[7] N. Cointe, G. Bonnet, and O. Boissier. Ethical judgment of agents' behaviors in multi-agent systems. In *15th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1106–1114, 2016.

[8] L. A. Dennis, M. M. Bentzen, F. Lindner, and M. Fisher. Verifiable machine ethics in changing contexts. In *35th AAAI Conference on Artificial Intelligence*, volume 35, pages 11470–11478. AAAI Press, 2021.

[9] G. Gervois, G. Bourgne, and M.-J. Lesot. Dealing with differentiated modalities of the Good : beyond Pareto optimality. In *International Workshop on AI Compliance Mechanism*, 2022.

[10] N. S. Govindarajulu and S. Bringsjord. On automating the doctrine of double effect. In *26th International Joint Conference on Artificial Intelligence*, pages 4722–4730, 2017.

[11] R. Limarga, M. Pagnucco, Y. Song, and A. Nayak. Non-monotonic reasoning for machine ethics with situation calculus. In *Advances in Artificial Intelligence*, pages 203–215, 2020.

[12] F. Lindner and M. M. Bentzen. A formalization of Kant's second formulation of the categorical imperative. In *14th International Conference on Deontic Logic and Normative Systems*, 2018.

[13] F. Lindner, M. M. Bentzen, and B. Nebel. The HERA approach to morally competent robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 6991–6997, 2017.

[14] C. Sarmiento, G. Bourgne, K. Inoue, D. Cavalli, and J.-G. Ganascia. Action languages based actual causality for computational ethics : a sound and complete implementation in ASP. *submitted*, 2023.

[15] C. Sarmiento, G. Bourgne, K. Inoue, and J.-G. Ganascia. Action languages based actual causality in decision making contexts. In *24th International Conference on Principles and Practice of Multi-Agent Systems*, 2022.

[16] Y. Taheri, G. Bourgne, and J.-G. Ganascia. A compliance mechanism for planning in privacy domain using policies. In *15th International Workshop on Juris-informatics*, 2021.

## De l'utilité de la réduction de la consommation énergétique des algorithmes d'intelligence artificielle

Par

**Paul GAY**
*LISN*
*Université Paris-Saclay*
*paul.gay@univ-pau.fr*

**Mohamed HEBIRI**
*LAMA*
*Université Gustave Eiffel*
*mohamed.hebiri@univ-eiffel.fr*

**Sébastien LOUSTAU**
*LMAP*
*Université de Pau et Pays de l'Adour*
*sebastien.loustau@univ-pau.fr*

**Florian VALADE**
*LAMA*
*Université Gustave Eiffel*
*florian.valade@univ-eiffel.fr*

### Introduction

L'intelligence artificielle (IA) est une discipline au croisement de l'informatique et des mathématiques qui s'attache à développer des machines capables de reproduire l'intelligence humaine en apprenant de leurs propres erreurs. Pour cela, cet apprentissage statistique résout un problème de minimisation de ce nombre d'erreurs à partir d'un grand nombre d'exemples potentiellement annotés par un humain. Dans le domaine de la santé, elle permet notamment de diagnostiquer des maladies, de proposer des traitements personnalisés et d'optimiser la gestion des hôpitaux. Dans l'industrie manufacturière, elle permet d'automatiser les chaînes de production, d'optimiser les processus de fabrication et de prévenir les pannes d'équipements. Elle est également une technologie clé des véhicules autonomes afin d'améliorer la sécurité routière et faciliter la mobilité. L'IA est enfin une composante majeure de la diffusion de l'information et de la publicité ciblée sur les réseaux sociaux. Cependant, la compréhension de son impact sociétal et environnemental peine à suivre son développement rapide. Outre la place grandissante qu'elle occupe aujourd'hui, l'IA questionne le futur et ses propres développements. L'impact environnemental n'étant que très partiellement mesurable, qu'est-ce qui justifie d'investir l'énergie de la recherche publique dans ce domaine en termes de retours positifs pour la société ?

## Empreinte environnementale de l'IA et du numérique

La majorité des travaux sur l'impact environnemental du numérique se concentrent principalement sur les émissions de gaz à effet de serre, pour lesquels le protocole de référence est le bilan carbone [6]. Tout d'abord, comme toute technologie numérique, on peut mesurer la consommation électrique de l'usage de l'IA, à la fois pour entraîner les modèles (étape d'*apprentissage* qui consiste à construire une règle de prédiction), et pour les utiliser en production (étape d'inférence). Le résultat dépendra du nombre d'opérations mathématiques effectuées (flops), mais aussi de la machine utilisée (*Graphic Processing Units* (GPU), *Central Proccessing Unit* (CPU)), ainsi que d'autres paramètres tels que la taille des données, voire les conditions extérieures comme température de la pièce, cette dernière influant sur la mise en place du ventilateur. Le tableau 1 donne quelques ordres de grandeur pour des modèles d'apprentissage profonds classiques sur des tâches de détection d'objet, classification d'images ou d'inférence sur du texte.

En considérant un mix énergétique aboutissant à l'émission de 60g de $CO_2$ par KWh, l'inférence de mille millions d'images avec le modèle de détection d'objet Yolov5s émettrait 10Kg de $CO_2$, soit l'équivalent d'environ 50km en voiture.

De plus, on peut ajouter à ces émissions liées à l'électricité utilisée par l'algorithme, les impacts liés à la fabrication et au recyclage des infrastructures et du matériel impliqué. Une telle étude, ayant répertorié l'ensemble des infrastructures supportant un modèle ou une application de l'IA n'existe pas à notre connaissance. En revanche, si l'on considère le secteur du numérique dans son ensemble [7], les observations suivantes peuvent être faites :

- En raison de leur grand nombre, les équipements numériques (appareils connectés excluant les *data centers*) concentrent 40 % des émissions de gaz à effet de serre du secteur pour leur fabrication, et 26 % de leur usage.
- Le trafic internet représente 16 % des émissions dont 80 % est composée de données vidéos.

Enfin, une méthodologie rigoureuse pour mesurer l'impact environnemental de l'IA doit inclure une analyse multi-critère, c'est-à-dire, qui ne se limite pas aux gaz à effet de serre, et ce, sur l'ensemble des cycles de vie [20] des composants et l'ensemble des usages des technologies de l'IA. Bien que la plupart des études se focalisent sur le bilan carbone, l'industrie du numérique est très liée à l'exploitation minière, les ordinateurs contenant plusieurs dizaines de métaux présents à l'état de traces, et dont l'extraction génère d'autres impacts, par exemple une grande consommation d'eau et une pollution des sols.

## Mesure de la consommation énergétique et méthodes de réduction

Depuis quelques années, de nombreux chercheur(e)s mesurent et proposent des méthodes pour réduire l'impact environnemental de l'IA [25].

### Mesure de l'énergie consommée directement par l'algorithme

Il existe des outils logiciels et physiques nous permettant de quantifier la quantité d'énergie utilisée par les algorithmes. En particulier, de nombreuses librairies open-source sont disponibles [2, 4, 5] pour mesurer la consommation d'énergie de chaque composant (processeur, différentes couches de mémoire, carte graphique, etc.) lors de l'exécution des programmes et des échanges de données. Ces outils sont tous basés sur (et donc limités par)

| Object Detector | CNN | Vision Transformer | Text Transformer |
|---|---|---|---|
| Yolov5s [8] | Resnet [14] | VIT_B_16 [12] | Bert [11] |
| 0.61 | 0.27 | 0.94 | 0.07 |

Table 1 – Joules consommés par un GPU sur une inférence. Voir [1] pour les détails expérimentaux (type de GPUs, taille du batch, etc.)

les moyens mis à disposition par les constructeurs tels qu'Intel ou Nvidia. Afin d'aller plus loin, la communauté du calcul distribué établit depuis plusieurs années des protocoles de mesure rigoureux avec les relevés physiques de wattmètres et en prenant en compte notamment l'usure des machines et le type de logiciel [22]. Il est courant qu'un coefficient de mix énergétique soit utilisé pour convertir l'énergie électrique consommée en équivalent $CO_2$. Ces outils permettent d'obtenir des ordres de grandeur et en général, la mesure permet de détecter certains gaspillages et mauvaises pratiques de programmation, par exemple la mauvaise utilisation du GPU. Cependant, malgré l'existence et la diffusion de ces outils, et la facilité de mise en œuvre, la communauté de l'IA a une forte tendance à résumer la qualité d'un modèle par rapport à un pourcentage de précision sur une tâche de référence. Par exemple, 10 années de recherche intensive d'une communauté de milliers de chercheur(e)s en vision par ordinateur (*computer vision*) ont permis d'augmenter la précision de classification d'images sur les données d'ImageNet, passant de 63 % à 91 % [3], sans considérer la consommation électrique ou l'impact environnemental de ces améliorations.

## Méthodes classiques de réduction d'énergie

Les techniques les plus populaires pour réduire le coût énergétique des algorithmes d'apprentissage profond sont : i) l'*optimisation des composants informatiques* en les dédiant à des tâches spécifiques – utilisation de GPU pour les tâches lourdes et de CPU dans le cas contraire ; ii) la *recherche d'architectures plus légères* qui consiste à simplifier l'encodage des poids appris par les réseaux [21] (*quantization*), à élaguer ces poids [9, 10, 15, 19] (*pruning*), à adapter l'algorithme d'inférence dynamiquement pour chaque donnée [13, 16, 17, 23], ou encore à distiller l'information de larges réseaux pré-entraînés vers de plus petits modèles [24]. Toutes ces stratégies exploitent l'hypothèse que les plus gros réseaux de neurones de l'état de l'art sont généralement calibrés pour la meilleure précision possible au détriment de leur complexité. De ce fait, des architectures réduites et performantes peuvent en être déduites grâce à l'apport des mathématiques (optimisation, statistiques, algèbre linéaire) pour obtenir des algorithmes qui intègrent directement les contraintes énergétiques. Ces méthodes sont prometteuses et peuvent permettre des gains significatifs dans l'entraînement des réseaux ou lors de l'inférence.

## Cas de l'inférence dynamique

Un cas particulier sur lequel nous travaillons est la sortie précoce où l'hypothèse est faite que certaines données sont plus faciles à analyser que d'autres et nécessitent donc moins de traitements pour arriver à une solution acceptable. Il est ainsi possible d'exploiter la structure par couches d'un réseau de neurones pour construire un ensemble de modèles à partir d'un seul. Plus précisément, la stratégie consiste à entraîner, au niveau de certaines couches ca-

chées du réseau, des modèles qui donneront lieu à des sorties précoces. L'utilisation de ces sorties précoces lors de la phase d'inférence permettra un gain énergétique intéressant par rapport à l'utilisation du modèle final. Ainsi, plus la sortie est précoce, moins l'énergie requise pour inférer est grande (Fig. 2).

Du point de vue théorique, cette approche est équivalente à l'estimation d'une règle optimale de rejet, et son implémentation requiert d'aller au-delà de l'aspect boite noire des réseaux de neurones afin de comprendre comment évolue l'information extraite au cours du processus d'inférence. Ce type d'approche conduit à des gains importants, par exemple, dans le traitement de vidéos où la majorité des images ne contiennent pas d'ambiguïté et doivent être rapidement analysées.

### Effet rebond ou paradoxe de Jevons

En pleine révolution industrielle, l'économiste néoclassique William Stanley Jevons remarquait le paradoxe suivant : alors que le progrès technique permettait aux chaudières à vapeur de consommer toujours moins de charbon, la consommation globale de charbon continuait de croître du fait de l'augmentation de leur nombre [18]. C'est le paradoxe de Jevons, ou *effet rebond*. Dans l'histoire récente de l'IA, l'introduction des GPU au début du XXIe siècle a permis d'accélérer significativement l'apprentissage des réseaux grâce à une parallélisation des calculs, et d'en réduire le coût énergétique. À titre d'exemple, un entraînement d'un modèle de reconnaissance vocale (*DeepSpeech*) sur GPU coûte 47kWh pour 150 heures de calcul. Ce même calcul prendrait environ 6 000 heures sur un processeur CPU, pour un coût énergétique 4 fois plus élevé. Cependant, l'introduction des GPU ne s'est pas accompagnée d'une réduction énergétique ! En effet, le gain en temps de calcul qui accompagne cette technologie est tel que de nouveaux algorithmes encore plus sophistiqués ont vu le jour, permettant d'améliorer significativement les performances et entraînant une augmentation de leur utilisation. Ainsi, la consommation d'électricité due à l'IA, loin d'avoir été divisée par 4, a considérablement augmenté. Aujourd'hui, de la même manière, nombre de travaux sur l'élaboration de sorties précoces et sur la réduction d'énergie présentés précédemment ont pour finalité de démocratiser l'IA à de nouveaux usages. On peut alors se questionner *in fine* sur l'impact environnemental de ces recherches.

## Bénéfices d'une réduction de la consommation de l'IA

En tant que piliers actuels de notre environnement numérique, et porteurs de nombreuses perspectives économiques et sociales, il est hautement probable que des algorithmes d'IA continueront à être utilisés dans les prochaines années et que de nouveaux usages apparaîtront. Dès lors, se pose la question de développer une IA à faible impact environnemental avec laquelle cohabiter.

Une partie de cet impact est due à la consommation énergétique associée aux machines effectuant les calculs. Cela peut avoir des conséquences importantes sur l'environnement, notamment en termes d'émissions de gaz à effet de serre et de dégradation des ressources naturelles. Réduire la consommation d'énergie des algorithmes d'IA via des algorithmes plus efficaces peut donc contribuer à atténuer ces impacts négatifs. De plus, l'analyse de l'efficacité des algorithmes permet de mieux comprendre le fonctionnement de modèles parfois qualifiés de boites noires.

En outre, en réduisant la consommation énergétique de l'inférence, les entreprises peuvent économiser sur les coûts de fonctionnement. La réduction des coûts liés à l'utilisation de l'IA est un enjeu important pour les entreprises et peut être réalisée à différents ni-

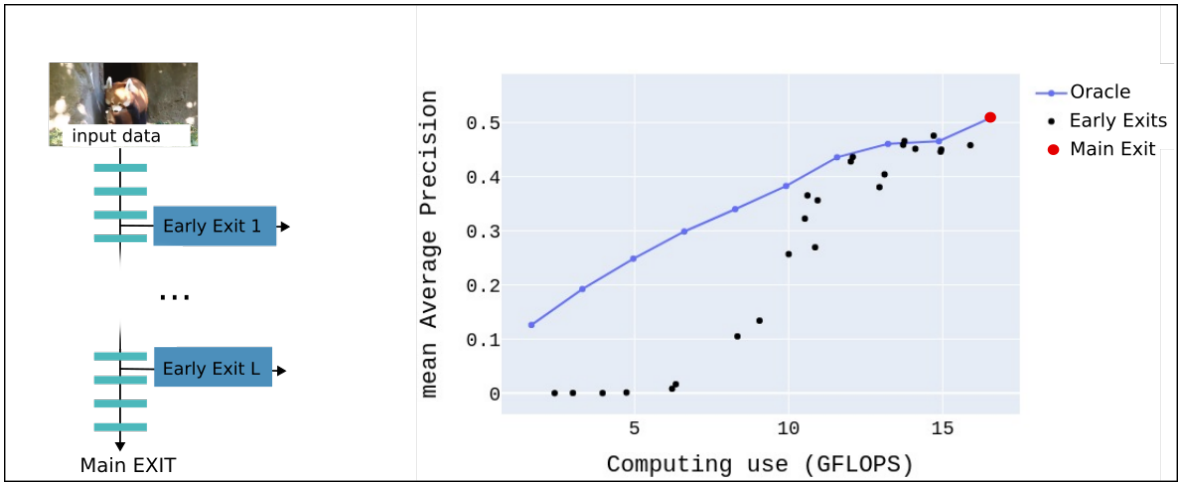



Figure 2 – *Gauche : illustration d'un réseau de neurones équipé de sorties précoces. Droite : précision et complexité de différentes sorties ajoutées sur un modèle de* deep learning *Yolov5 sur les données COCO avec un oracle sélectionnant les meilleures sorties précoces pour maximiser la précision étant donné une contrainte de budget énergétique et permettant d'explorer différents compromis précision/consommation.*

veaux, de l'entraînement à l'inférence. La réduction de la consommation énergétique est un moyen de réaliser des économies tout en contribuant à la préservation de l'environnement.

Enfin, la réduction des coûts énergétiques permettrait de déployer des réseaux de neurones plus vastes ou des architectures plus complexes. Elle passe aussi bien souvent par une amélioration du temps de latence, ouvrant la voie à de nouveaux usages comme les applications temps réel ou dans le domaine de l'embarqué. Ce dernier point peut cependant être contre productif du point de vue environnemental en générant de nouveau un effet rebond, comme décrit dans la section précédente.

## Conclusion

L'impact environnemental de l'intelligence artificielle dû à l'explosion de l'utilisation de cette technologie, et à des algorithmes et des quantités de données toujours plus importantes, est un enjeu majeur pour l'avenir. Dans ce domaine, la communauté scientifique en mathématiques et en informatique s'intéresse depuis plusieurs années à la mesure et à la réduction de la consommation électrique des algorithmes. Les limites de ces travaux sont nombreuses et devront être explorées dans le futur. Tout d'abord, les ordres de grandeur nous montrent que la consommation des algorithmes n'est qu'une petite partie du problème. Par exemple, les émissions de carbone liées au déplacement d'un chercheur parcourant 25 km en voiture pour aller à son bureau pendant deux ans correspondent à la consommation d'un modèle Yolov5 appliqué à environ 4 millions d'heures de vidéo [31]. De plus, la consommation relative à l'utilisation des algorithmes n'est qu'une partie de l'impact direct de ces technologies, auquel il faut ajouter l'énergie grise utilisée pour produire les appareils et les infrastructures de calculs, à l'échelle du centre de calcul, ou du composant embarqué. Enfin, pour permettre une analyse sérieuse

31. La durée cumulée que les Français passent sur Youtube en une journée est du même ordre de grandeur.

de l'impact environnemental des technologies de l'IA, l'usager final de chaque algorithme doit être pris en compte, pour permettre un bilan objectif. Quels sont ces objectifs et quelles sont les conséquences en termes d'émissions de $CO_2$, d'exploitation de ressources, ou plus largement de pression écologique ?

## Références

[1] AiPowerMeter library. Deep learning benchmark. Accessed : 2023-03-11.

[2] Codecarbon : track and reduce $co_2$ emissions from your computing. Github. Accessed : 2023-03-10.

[3] Image classification on ImageNet. Papers with Code. Accessed : 2023-03-10.

[4] Scaphandre. Github. Accessed : 2023-03-10.

[5] L. F. W. Anthony, B. Kanding, and R. Selvan. Carbontracker : Tracking and predicting the carbon footprint of training deep learning models. *arXiv preprint :2007.03051*, 2020.

[6] Association bas carbone. Guide méthodologiques V8.7 Bilan Carbone, Objectifs et principes de comptabilisation. https://abc-transitionbascarbone.fr/, February 2023. Accessed : 2023-02-02.

[7] F. Bordage. Empreinte environnementale du numérique mondial. *Green IT*, page 9, 2019.

[8] Ultralytics Company. Yolov5, 2021. Accessed : 2023-02-02.

[9] Y. Cun, J. Denker, and S. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems 2*, pages 598–605. 1990.

[10] P. de Jorge, A. Sanyal, H. Behl, P. Torr, G. Rogez, and P. Dokania. Progressive skeletonization : Trimming more fat from a network at initialization. In *9th International Conference on Learning Representations*, 2021. arXiv : 2006.09081.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint : 1810.04805*, 2018.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words : Transformers for image recognition at scale. *arXiv preprint : 2010.11929*, 2020.

[13] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang. Dynamic neural networks : A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11) :7436–7456, 2021.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[15] M. Hebiri and J. Lederer. Layer sparsity in neural networks. *arXiv preprint : 2006.15604*, 2020.

[16] Y. Kaya, S. Hong, and T. Dumitras. Shallow-deep networks : Understanding and mitigating network overthinking. In *36th International Conference on Machine Learning*, pages 3301–3310, 2019.

[17] S. Laskaridis, A. Kouris, and N. Lane. Adaptive inference through early-exit networks : Design, challenges and directions. In *5th International Workshop on Embedded and Mobile Deep Learning*, pages 1–6, 2021.

[18] S. Latouche. *Le pari de la décroissance*. Fayard, 2006.

[19] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang. Pruning and quantization for deep neural network acceleration : A survey. *Neurocomputing*, 461 :370–403, 2021.

[20] Negaoctet pour ScoreLCA. Impacts environnementaux des objets connectés et des services basés sur leur utilisation : Ordres de grandeurs et recommandations méthodologiques, 2021. Accessed : 2023-03-10.

[21] P.-E. Novac, G. Hacene, A. Pegatoquet, B. Miramond, and V. Gripon. Quantization and deployment of deep neural networks on microcontrollers. *Sensors*, 21(9) :2984, 2021.

[22] A.-C. Orgerie. *From Understanding to Greening the Energy Consumption of Distributed Systems*. PhD thesis, École Normale Supéieure de Rennes, 2020.

[23] S. Teerapittayanon, B. McDanel, and H.-T. Kung. Branchynet : Fast inference via early exiting from deep neural networks. In *23rd International Conference on Pattern Recognition*, pages 2464–2469, 2016.

[24] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. In *38th International Conference on Machine Learning*, pages 10347–10357, 2021.

[25] R. Verdecchia, J. Sallou, and L. Cruz. A systematic review of green AI. *arXiv preprint : 2301.11047*, 2023.

# La Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires

Par **Christian BYK**
*Magistrat honoraire*
*Chercheur associé, Institut des sciences juridique et philosophique de la Sorbonne*
*Président du Comité intergouvernemental de bioéthique de l'UNESCO (2017 – 2019)*
*christian.byk@gmail.com*

## Introduction

Le droit et le système de production normative, dont la justice est l'une des composantes, apportent une contribution importante et nécessaire à l'organisation de nos sociétés. Mais, la justice est aussi affectée par ce qui caractérise aujourd'hui nos sociétés : la mondialisation, la complexification, le poids de l'économie et de la technoscience ainsi que l'individualisme. Accessible à un plus grand nombre de justiciables depuis les années 1970, le système judiciaire fait ainsi face à un nouveau défi : celui d'une gestion des flux rationalisée. La révolution numérique peut-elle alors, comme la mécanisation pour la dentelle ou le taylorisme pour l'automobile, apporter à chacun ce qu'il estime être son dû : plus d'égalité, plus de rapidité et de sécurité juridique à un moindre coût ? Est-ce possible sans « déshumaniser » la justice ? Les applications de l'IA dans le domaine de la justice connaissent depuis les années 2010 un développement fulgurant qui a fait émerger de nouveaux acteurs, les *legaltech*, de sorte que les analystes s'accordent pour dire que « c'est l'émergence de la *legaltech* sur la scène internationale et nationale qui impose aujourd'hui avec une certaine urgence une réflexion plus approfondie sur les conséquences de l'usage crois-

sant des technologies numériques dans le secteur de la justice »[2]. S'il existe donc un avenir prometteur pour les *legaltech*, en termes de marché tout autant que de meilleure effectivité du système judiciaire, comment cet avenir influencera-t-il la transformation, voire la réformation de la justice ? Les attentes sociales à cet égard étant grandes, on peut penser que les *legaltech* devraient faire entrer la justice dans le XXIe siècle, celui d'une efficacité interactive. Dans cette perspective, quatre fonctions principales de l'IA sont en susceptibles d'être développées dans le domaine de la justice :

- *générer de la connaissance* en mettant en évidence les manières différenciées de rendre la justice selon les régions,
- *prédire* la chance de succès d'un procès et le montant potentiel des dommages-intérêts,
- *recommander* des solutions de médiation en fonction du profil des personnes et des cas similaires passés,
- *aider à la décision* en suggérant au juge la solution jurisprudentielle la plus adéquate pour un cas.

Le fil conducteur de notre réflexion ne doit-il pas alors se concentrer dans la question : le numérique est-il un outil pour servir ou s'approprier la justice ?

## Une réponse normative : la Charte éthique européenne sur l'utilisation de l'intelligence artificielle dans les systèmes judiciaires et leur environnement

En adoptant, en 2018, « la Charte éthique européenne sur l'utilisation de l'intelligence artificielle dans les systèmes judiciaires et leur environnement » [1], le Conseil de l'Europe nous donne une première réponse normative. Certes, celle-ci ne remet pas en cause la pertinence à utiliser l'outil de l'IA, mais elle en rappelle le cadre : « Le respect des droits fondamentaux des individus énoncés par la Convention européenne des droits de l'homme (CEDH) et la Convention n° 108 du Conseil de l'Europe pour la protection des données à caractère personnel. » La Charte revêt néanmoins une importance non négligeable. D'une part, elle s'adresse à 46 États membres, qui reconnaissent tous la compétence de la Cour européenne des droits de l'homme, mais aussi aux entreprises privées aussi bien qu'aux acteurs publics. D'autre part, son approche méthodologique se veut concrète : à l'affirmation de principes s'ajoutent non seulement une étude approfondie sur les applications de l'IA dans les systèmes judiciaires ainsi qu'une classification des utilisations suivant qu'elles sont ou non à encourager, mais encore une *checklist* d'intégration des principes de la Charte dans les traitements automatisés.

La Charte met ainsi en avant cinq principes :

- le principe de respect des droits fondamentaux : il s'agit d'intégrer dès la phase de conception des règles interdisant de porter atteinte aux droits fondamentaux,
- le principe de non-discrimination : les acteurs publics doivent s'assurer que les traitements ne reproduisent pas ou n'aggravent pas les discriminations et qu'ils ne conduisent pas à des analyses ou usages déterministes,
- le principe de qualité et de sécurité : les données dérivant des décisions juridictionnelles et intégrées dans un logiciel qui exécute un algorithme d'apprentissage-machine doivent provenir de sources certifiées et ne doivent pas pouvoir être altérées,
- le principe de transparence, de neutralité et d'intégrité intellectuelle, qui impose qu'« un équilibre doit être trouvé entre la propriété intellectuelle de certaines méthodes de traitement et les exigences de transparence , de neutralité (absence de biais), de loyauté et d'intégrité intellectuelle »,

- le principe de maîtrise par l'utilisateur : le juge doit pouvoir revenir aux décisions et données judiciaires utilisées et continuer à avoir la possibilité de s'en écarter tandis que le justiciable devrait être informé dans un langage clair et compréhensible du caractère contraignant ou non des solutions proposées par les outils d'intelligence artificielle.

L'annexe II est particulièrement opportune pour les praticiens, car elle passe en revue différentes utilisations de l'IA et encourage à un degré différent leur application à la lumière des principes et des valeurs énoncés dans la Charte. Si la valorisation de la jurisprudence et l'accès au droit sont encouragés « sans réserve », en revanche, l'aide à la construction de barèmes, l'appui à des mesures alternatives de règlement de litiges en matière civile, le règlement des litiges en ligne et l'utilisation des algorithmes en matière d'enquête pénale afin d'identifier des lieux de commission d'infractions ne peuvent être encouragés qu'avec de fortes précautions méthodologiques. Enfin, le profilage des juges, l'anticipation des décisions supposent des travaux scientifiques complémentaires, tandis que l'utilisation des algorithmes en matière pénale afin de profiler les individus et le fait d'enfermer le choix du juge dans la masse des « précédents » ne peuvent être envisagés qu'avec les plus extrêmes réserves.

## Conclusion

Indicative plus que prescriptive, la Charte suppose une implication dynamique de tous les acteurs pour devenir un instrument vivant. Mais s'ils manquaient à répondre à cette invitation, ces acteurs devraient assumer la responsabilité d'une perte de confiance des citoyens dans la capacité de la justice à respecter les valeurs d'une société démocratique.

## Références

[1] Commission européenne pour l'efficacité de la justice (CEPEJ). Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires et leur environnement, Conseil de l'Europe, Note de présentation, Strasbourg, 3-4 décembre 2018.

[2] L. D. Godefroy, F. Lebaron, and J. Lévy-Vehel. *Comment le numérique transforme le droit et la justice vers de nouveaux usages et un bouleversement de la prise de décision*. Rapport de recherche, Mission de recherche Droit et Justice, ffhal-02264192, 2019.

# L'Ingénierie de la Connaissance à l'heure de l'ISO30401

Par **Alain BERGER**
*Ardans SAS*
*aberger@ardans.fr*
*www.ardans.fr*

## De la connaissance dans l'IA jusqu'à la norme ISO30401

Il est toujours très intéressant de positionner l'*Ingénierie de la Connaissance* dans les différentes disciplines qui constituent l'intelligence artificielle. En partant de la conjecture de Dartmouth proposée en 1955, postulat selon lequel « chaque aspect de l'apprentissage ou toute autre caractéristique de l'intelligence peut être si précisément décrit qu'une machine peut être conçue pour le simuler » [8], la problé-

matique de l'intelligence artificielle était posée comme « une tentative . . . faite pour trouver comment faire en sorte que machines utilisent le langage, forment des abstractions et des concepts, à résoudre des types de problèmes aujourd'hui réservés aux humains, et à s'améliorer d'elles-mêmes ».

Près de soixante-dix ans plus tard, avec les progrès indéniables des techniques de l'apprentissage et des agents conversationnels, il est particulièrement intéressant d'observer les évolutions de cette discipline centrée sur les connaissances depuis les *systèmes experts* vers celle du *Management de la Connaissance* (ou *Knowledge Management*). Comment depuis la mise en œuvre de moteurs de règles, les réflexions sur leur implantation opérationnelle au sein d'organisation humaine, les apports de Ikujiro Nonaka [9] à Michel Grundstein [6] se retrouvent avec un certain niveau d'agrégation dans la norme ISO30401 [11] ?

## L'ingénierie dédiée à la connaissance

La modélisation de la connaissance est un sujet qui depuis les années quatre-vingts continue d'évoluer au fil des nouveaux formalismes et langages proposés aux ingénieurs.

Nous avons sélectionné deux schémas parmi ceux proposés par ChatGPT-4 le 17 mars 2023 (cf. figure 3) en réponse à la question « Aurais-tu un schéma qui représente les techniques de modélisation en ingénierie des connaissances ? ». Au-delà de la magie de réponses immédiates et intéressantes, l'usager est en droit de s'interroger sur la variabilité, voire la consistance des réponses produites. Obtenir *n* réponses distinctes en réitérant la même requête génère un trouble ; la situation est ici inhérente à l'algorithme sous-jacent.

Entre les *frames*, les systèmes experts et leurs règles, la logique formelle et ses prédicats, les objets, les *blackboards*, les *truth maintenance systems*, les méta-connaissances, les réseaux bayésiens, les graphes de connaissances, les ontologies, les modèles de Markov, etc., à chacun son formalisme où deux objectifs étaient concourants, voire concurrents : optimiser les performances et fournir la bonne réponse.

Mais si, initialement, il y avait une compétition sur les « moteurs », avec le temps celle-ci s'est déplacée vers la question de la qualité de la production et de l'accès à la connaissance.

## La production de la connaissance

Comment la connaissance se fabrique-t-elle ? Léonard de Vinci précisait que « toutes nos connaissances ont pour origine notre perception ». Oui, certes pour la genèse, il y a une sorte d'intuition, mais la question de la preuve est essentielle, comme le pressent Platon : « la connaissance est une croyance vraie et justifiée ». Nous n'évoquerons pas ici la question posée par Edmund Gettier qui s'interroge sur le fait que cela soit nécessaire et suffisant.

Il n'en reste pas moins que l'obsession de la justification, du calcul, de la démonstration, de la preuve anime sans relâche le scientifique. Le contexte et le processus qui conduisent à la connaissance sont essentiels, ainsi que le rappelle Étienne Klein : « Savoir, c'est en somme savoir *comment on a su*. »

Ces éléments se révèlent comme fondateurs à la compréhension et surtout à la confiance dans le résultat délivré à l'utilisateur !

## L'élicitation de la connaissance vivante

Comme ce que l'on appelle *connaissance* est aussi le fruit de l'expérience, elle n'a pas été forcément formalisée dans un cadre qui soit un dispositif de restitution de connaissance : elle est en tous les cas portée par l'humain au sein de son cerveau. Le travail de l'ingénieur de la connaissance est de découvrir cette pépite et de la révéler, de l'extraire pour la mettre

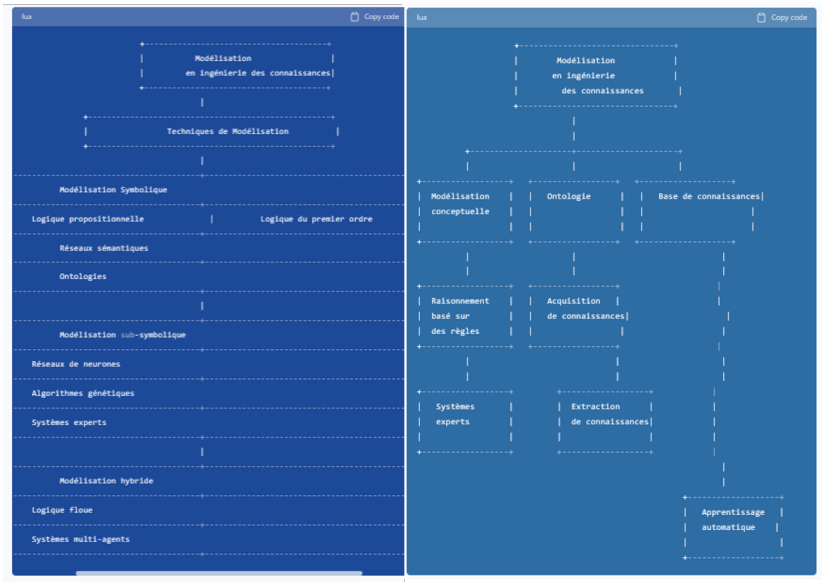



Figure 3 – Schémas des techniques de modélisation en IC selon ChatGPTv4

dans la forme la plus fidèle et conforme à ce que le sachant a pu exprimer lors des entretiens. En suscitant, en stimulant, en provoquant l'expert, l'ingénieur de la connaissance tire de l'expérience du sachant, lui fait sortir son savoir : il s'agit du processus d'élicitation de la connaissance. Cette démarche est très importante quand il s'agit de mettre à nu de la connaissance *implicite*, ce que soulève Jean-Yves Prax lorsqu'il mentionne le cas d'un expert surpris et s'exclamant : « on ne sait pas ce qu'on sait ! » [10].

Pour être plébiscités dans l'industrie, la base de connaissance ou le *Système de Management de la Connaissance* (SKM) doivent satisfaire aux attentes des acteurs dont :

- ▷ « *exhaustivité* » : il convient que la connaissance soit exhaustive sur le périmètre sur lequel elle est porte afin d'obtenir la confiance de l'utilisateur à commencer par lui transmettre la réponse pertinente ;
- ▷ « *consistance* » : les résultats de « navigation » pour obtenir les contenus sont consistants ; cette stabilité rassure l'utilisateur ;
- ▷ « *clarté* » : les contenus sont clairs, dénués de toute ambiguïté, cela pour faciliter l'adhésion, l'appropriation et le bon usage par l'utilisateur ;
- ▷ « *argumentation* » : les contenus sont argumentés et disposent des niveaux de preuve nécessaires pour une bonne appropriation par le lecteur ;
- ▷ « *contextualisation* » : il est fondamental de bien décrire le contexte dans lequel cette connaissance est valide pour être exploitée en toute sérénité ;
- ▷ « *position* » : l'élément de connaissance consulté est au cœur d'un réseau (implicitement sémantique) d'éléments de connaissance au sein desquels il doit être positionné dans une représentation cartographique multidimensionnelle : un réseau précieux pour évaluer la qualité de la base comme son homogénéité, ses relations, ses trous, ses densités ;
- ▷ « *diffusion* » : la connaissance est un actif précieux et est restreinte à ceux habilités à en connaître, celui qui en bénéficie doit sa-

voir le mesurer ;

- « *convivialité* » : plus que jamais l'ergonomie d'un système à base de connaissance moderne doit être d'une ergonomie intuitive et fluide et démontrer qu'elle offre un retour sur investissement à l'usager sans pareil.

Pour l'organisme, le SKM a indubitablement la qualité de réaliser le *transfert de connaissance* vers l'utilisateur tel que défini par l'équation de Davenport et Prusak [4] : *Transfer = Transmission + Absorption (and use)*.

Ce qui est certain, c'est qu'aujourd'hui les outils de modélisation de connaissance disposent d'environnements de visualisation (cf. l'exemple du graphe en figure 4) comme de fonctionnalités d'agrégation extrêmement riches pour distiller la bonne connaissance au bon moment à l'utilisateur (cf. l'exemple de l'élément pointé par Parnasse en figure 6).

## L'implantation de la connaissance colligée

Si comme l'exprime François Vexler [12] « la connaissance, cela se mérite ! », en expert de l'*Ingénierie de la Connaissance* , il sait que la clé d'un dispositif pérenne est de réaliser la bonne modularisation et structuration de la connaissance pour une exploitation vertueuse et fructueuse.

Il s'agit de trouver le juste équilibre entre une modularisation trop fine qui rebutera le futur contributeur pour alimenter et faire vivre dans le temps la base de connaissance, et une modularisation trop grossière qui ne guidera pas le futur lecteur pour trouver la réponse à sa question métier.

Le choix des modèles est d'autant plus aisé qu'il va coller à un processus métier, à une cinématique opérationnelle, fussent-ils complexes.

La structuration, par ailleurs, s'impose en s'appuyant sur un langage métier partagé et une ontologie dénuée par construction de toute ambiguïté pour classer les concepts. Ces règles édictées [2, 7], la question de la conduite du changement et de la transmission du système à la maîtrise d'ouvrage devient prioritaire et c'est là que le risque se transfère vers la partie culturelle comme organisationnelle pour la dissémination de la démarche.

## L'Ingénierie de la Connaissance

Actualiser la définition proposée en 2013 [1] est nécessaire.

L'*Ingénierie de la Connaissance* est une discipline de l'IA qui couvre tout un cycle, depuis l'*élicitation* d'un élément de connaissance, sa *structuration*, son *mûrissement* en termes de contenu, sa *description* (via une définition claire, non ambiguë, la rédaction étant appuyée par des illustrations ou schématisations si nécessaire), son *applicabilité* (en termes de domaine d'usage, de droit à en connaître en termes de publication ou d'habilitation, de durée de vie ou de date de péremption), et bien sûr de *validation* (appréciation d'expert, justification, degré de preuve).

Lorsque l'on travaille sur la mémoire collective d'un domaine métier, il convient d'orchestrer les différents *modèles* qui représenteront les éléments de connaissances, de poser les *liens de sémantique* entre les éléments en relation, et les *liens de classification* de ces éléments par rapport à des*ontologies* descriptives des concepts métier partagées, intelligibles, distinctes, complémentaires et non contestables.

L'*Ingénierie de la Connaissance* offre à l'utilisateur « lecteur » les moyens de trouver la connaissance auquel il aspire et à se l'approprier, à l'utilisateur « contributeur » la capacité à actualiser le patrimoine auquel il a accès, à l'utilisateur « gestionnaire » pour le compte de l'organisme, la capacité à exploiter son capital connaissance selon les règles de dissémination ou de protection en conformité avec son attente ou avec la réglementation.

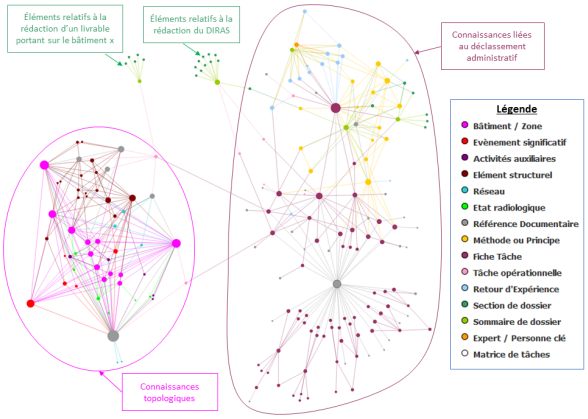



Figure 4 – La cartographie d'une base en construction

## Le management des connaissances et l'ISO30401

### La publication de la norme ISO30401

En novembre 2018, la norme ISO30401 intitulée « Systèmes de management des connaissances – Exigences » [11] est publiée par le Comité technique Management des Ressources Humaines de l'Organisation Internationale de Normalisation (ISO). Sa finalité est « d'aider les organismes à concevoir un système de management qui valorise et facilite la création de valeur grâce aux connaissances ».

Ce qui est remarquable, c'est que cette norme s'attaque de manière exhaustive à toutes les facettes du *Management de la Connaissance*.

### La première analyse de l'ISO30401

◇ Le premier résultat offert à tous par l'arrivée de cette norme est ce *vocabulaire commun* partagé par toute une communauté de professionnels de la discipline. Cette fondation est très importante quand on connaît tout l'éventail de provenance des contributeurs : culture, langue, métier, nature, taille et ressources de l'organisme.

◇ Le deuxième avantage est le fait que les trois grands fondamentaux classiques du *Management de la Connaissance*se retrouvent dans la norme (cf. le schéma de la reformulation par Ardans figure 5), à savoir :

1. La modélisation **SECI** de Nonaka & Takeuchi[9] (SECI pour Socialisation, Externalisation, Combinaison, Internalisation) avec le double positionnement Implicite/Explicite et Individuel/Collectif,
2. Les cinq facettes de la *problématique de capitalisation des connaissances* de Grundstein [5] (Repérer, Préserver, Valoriser, Actualiser, Manager),
3. La roue de Deming **PDCA** (pour *Plan, Do, Check, Act* ou Planifier, Réaliser, Vérifier, Agir) relative à l'amélioration continue, et accompagnée par tout un arsenal d'éléments facilitateurs (humains, processus, technologies, gouvernance, culture).

◇ Le troisième point important est la vision de **S**ystème de **M**anagement de la **C**onnaissance (SMC ou SKM) avec toutes les dimensions de l'ingénierie système appliquée au KM comme l'indique Patrick Coustillière[3], pour être synthétique, ce qui est relatif aux **E**xigences, **R**essources, **O**rganisation [Rôle], **P**rocessus, **E**xports [Fournitures].

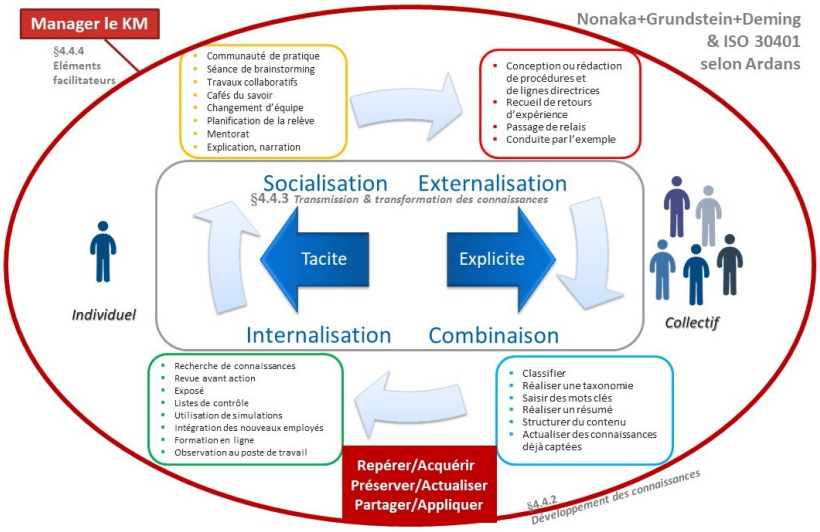



Figure 5 – Reformulation de l'ISO30401 selon Ardans

◇ Le quatrième point est astucieux car il délivre une clé pour les organismes en difficulté pour traiter le « *§7.1.6 Connaissances organisationnelles* » de la norme ISO9001:2015 relative aux systèmes de management de la qualité, ce chapitre qui détient la palme de première non-conformité lors des audits. En précisant, qu'il s'agit d'une clé, nous entendons que si l'on trouve dans la norme les ingrédients, la recette n'y figure pas.

## Parnasse : vers une meilleure appropriation de la norme

Si dans certains pays, des consultants se sont proclamés « auditeurs » de la norme ISO30401, en Europe la prudence est de rigueur. Pour autant, le Club Gestion des connaissances [32] a souhaité valoriser ses travaux internes autour de son « SKM Book » et le rendre plus accessible en tant que « SKM de référence » qui satisfait aux exigences de l'ISO30401. C'est ainsi que Parnasse [33] a été conçu comme « l'atelier du knowledge manager qui l'aide à évaluer et à améliorer leur système KM au regard de la norme ISO30401 » (cf. figure 6).

L'objectif est de rendre possible l'exploitation « multi-vues » de ce référentiel complexe grâce à la puissance de modélisation de l'outil sélectionné : Ardans Knowledge Maker®.

La navigation dans la version logicielle Parnasse illustre clairement les vues du SKM Book. L'utilisateur sait ainsi :

- Quelles activités du SKM Book afin de satisfaire aux exigences de la norme ?
- Quelles exigences de la norme sont concernées par les activités du SKM Book ?
- Pour une action KM, quelles activités du SKM Book sont préconisées et quelles exigences de la norme sont concernées ?

Avec Parnasse, le *Knowledge Manager* sait évaluer les actions KM déjà en place, les situer au sein d'un SKM référence et ainsi identifier un plan de route vers une cible plus complète en accord avec la norme.

L'expert et concepteur de Parnasse, Da-

32. Association loi de 1901 créée en 2000.
33. **P**ortail **A**rticulant la **R**éférence **N**ormative iso30401 **A**vec un **S**ystème KM **S**tructuré pour l'**E**ntreprise.

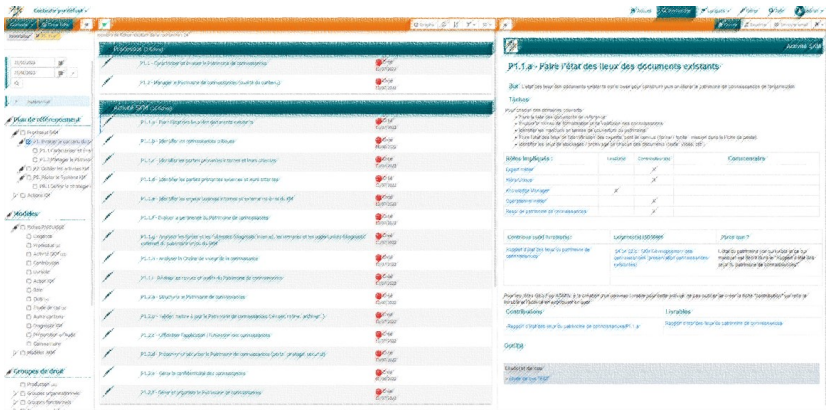


Figure 6 – Parnasse : mieux appréhender son SKM selon l'ISO30401 via Ardans Knowledge Maker®

niel Colas, considère qu'en donnant les clés et réponses à la compréhension du référentiel ISO30401 : « Parnasse *démontre qu'une étape majeure de maturité dans la discipline est atteinte !* »

## Perspectives

La question de la normalisation dans le domaine de l'intelligence artificielle est un sujet qui peut être traité par sous-discipline ainsi que le démontre cette norme ISO30401 dédiée aux *systèmes de management des connaissances*.

Pour autant, il convient d'être prudent, car si la norme a l'avantage de mettre à plat le langage commun des référents pour la discipline, elle ne donne aucune indication sur le comment faire, ni sur les pièges à éviter.

On note que des associations ou sociétés savantes référentes dans un domaine de l'IA sont aussi parfaitement capables de faire émerger des outils d'accompagnement à la compréhension fine de la norme. L'exemplarité de Parnasse co-produit par le Club Gestion des Connaissances et Ardans en est la preuve pour le domaine de l'*Ingénierie de la Connaissance*et le *Management de la Connaissance*.

Comme le disait Alfred Korybski « *la carte n'est pas le territoire !* » ; si la norme peut couvrir l'ensemble des activités qui couvrent l'*Ingénierie de la Connaissance*, ce métier est extrêmement humain dans l'activité d'élicitation de la connaissance. En conséquence, des résultats toujours différents seront produits en fonction des professionnels qui l'exerceront.

La maîtrise technique de cette discipline reste aujourd'hui un art, dans la relation humaine, comme dans la transcription vers le système pour garantir la meilleure exploitation future en toute confiance.

## Références

[1] A. Berger. L'ingénierie de la connaissance et la mémoire collective au cœur de la dynamique éthique des organisations. *Bulletin de l'AfIA*, 79 :13–16, 2013.

[2] V. Besson and A. Berger. To initiate a corporate memory with a knowledge compendium : ten years of learning from experience with the Ardans method. In *15es Journées Francophones Extraction et Gestion des Connaissances*, pages 401–412, 2015.

[3] P. Coustillière. L'ingénierie système, un outil pour le km manager ? Revues des Nouvelles Technologies de l'Information, Mai 2022.

[4] T. Davenport and L. Prusak. *Working Knowledge : How Organizations Manage what They Know*. EBSCO eBook Collection. Harvard Business School Press, 1998.

[5] M. Grundstein. Développer un système à base de connaissance : un effort de coopération pour construire en commun un objet inconnu. In *Acte de la journée Innovation pour le travail en groupe*, 1994.

[6] M. Grundstein. CORPUS, an approach to capitalizing company knowledge. In *4th International Workshop on Artificial Intelligence in Economics and Management*, 1996.

[7] P. Mariot, C. Golbreich, J.-P. Cotton, and A. Berger. Méthode, modèle et outil Ardans de capitalisation des connaissances. In *7es Journées Francophones Extraction et Gestion des Connaissances*, pages 187–206, 2007.

[8] J. McCarthy, M. Minsky, N. Rochester, and C. Shannon. A proposal For the Dartmouth summer research project on Artificial Intelligence, 1955.

[9] I. Nonaka and T. Hirotaka. *The Knowledge-Creating Company : How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press, 1995.

[10] J.-Y. Prax. *Manuel de Knowledge Management - 4e édition*. Les Actus du Savoir : Management/Leadership. Dunod, 2019.

[11] ISO Central Secretary. Knowledge management systems – Requirements ISO30401:2018. 2018.

[12] F. Vexler, A. Berger, J.-P. Cotton, and A. Belloni. Eléments d'appréciation et d'analyse d'une base de connaissance : l'expérience industrielle d'Ardans. In *Atelier AIDE à la 13e Conférence Francophone sur l'Extraction et la Gestion des Connaissances*, pages 59–72, 2013.

# Non-prolifération de l'intelligence artificielle générative – Quel encadrement normatif ?

Par **Eva THELISSON**
*AI Transparency Institute*
*eva@aitransparencyinstitute.com*
*www.aitransparencyinstitute.com*

« Où est la vie que nous avons perdue en vivant ? Où est la sagesse que nous avons perdue dans la connaissance ? Où est la connaissance que nous avons perdue dans l'information ? »
*T. S. Eliot, Chorus from The Rock.*

## Introduction

Cet article aborde la question de la prolifération des systèmes d'intelligence artificielle (IA) tant dans le domaine civil que militaire. Il discute du besoin éventuel d'une réglementation spécifique au niveau international pour encadrer l'usage des systèmes d'IA générative (en anglais *General Purpose AI* (ci-après GPAI)). Il présente les défis posés par les systèmes GPAI et les efforts de réglementation en cours. Du fait des possibilités de double usage des systèmes GPAI, ce papier soulève la question d'un encadrement normatif pour faciliter le contrôle de ces systèmes au niveau international. Dans

ce contexte, il propose une réflexion comparée avec le droit international nucléaire.

## Définitions

La notion de prolifération se réfère à une multiplication rapide[34]. Dans le domaine nucléaire, le dictionnaire Larousse précise qu'il s'agit également de l'augmentation du nombre des nations accédant à une capacité nucléaire militaire indépendante. *A contrario*, le concept de non-prolifération se définit comme la limitation de la production et du stockage des armes nucléaires dans le monde.

La disponibilité de grandes quantités de données et de puissances de calcul accélère la dissémination rapide d'algorithmes d'apprentissage automatique capables de traiter et d'apprendre à partir de grandes quantités de données. Les systèmes à base d'IA sont capables de s'adapter en fonction de leur expérience, ce qui leur permet d'effectuer des tâches qui requièrent généralement l'intelligence humaine, telles que la perception, la prise de décision et la compréhension du langage. Plusieurs domaines de recherche ont bénéficié de ces avancées majeures, par exemple, la vision par ordinateur, le traitement du langage naturel et la reconnaissance vocale. Les systèmes d'IA générative reposent quant à eux sur des réseaux de neurones artificiels, qui sont capables de produire des données sur la base de données initiales (figure 7).

## Pour une vérification des GPAI

La lettre ouverte du *Future of Life Institute* (FLI), signée en mars 2023 par plus de 3000 spécialistes en IA [11] alerte sur la prolifération des systèmes d'IA générative. Elle appelle les laboratoires d'IA à effectuer une pause dans le développement des systèmes plus puissants que chatGPT-4. Les signataires demandent plus précisément « la vérification et la transparence de ces systèmes ». A défaut, les signataires réclament un moratoire de la part des gouvernements.

## Une course aux armements

Depuis la sortie de ChatGPT en novembre 2022, Microsoft a investi 10 milliards de dollars pour obtenir la priorité de l'usage de cet outil conversationnel [6]. Google et Meta ont décidé de concevoir leur propre IA générative : Google a conçu Bard sur la base du modèle de langage LaMDA et Meta a dévoilé LLaMA, un modèle de langage avec moins de paramètres que GPT-3 pour soutenir la communauté de recherche. Amazon Web Service vient d'annoncer en avril 2023 la création d'un accélérateur dédié aux startups spécialisées dans l'IA générative avec plus de 300 000 dollars de dotation par projet.

## Un enjeu géopolitique

Cette course aux armements dans le domaine de l'IA générative est à replacer dans le contexte des tensions internationales dans le domaine de la production de semi-conducteurs et de la fabrication de puces électroniques, qui ont donné lieu aux *Chips Act* américain, coréen et européen. Selon Yasutoshi Nishimura, ministre japonais de commerce, « les puces de pointe (2 nanomètres), seront essentielles pour un large éventail d'industries dans les cinq à dix prochaines années, telles que les services d'intelligence artificielle de type ChatGPT ainsi que l'informatique quantique » [12]. Contrôler les applications de l'IA générative, c'est assurer un avantage compétitif, tant dans les domaines industriels, commerciaux, de sécurité et les domaines militaires. Ralentir le marché représente donc un risque de perte de compétitivité dans un monde multipolaire concurrentiel [10].

34. Dictionnaire Larousse, V° Prolifération


Figure 7 – Exemples d'applications d'IA générative (source Zain Kahn)

## Systèmes GPAI : Quelle réglementation ?

De manière similaire au droit nucléaire, qui tente d'empêcher l'émergence d'une capacité d'armement nucléaire en empêchant le développement d'un processus d'enrichissement de l'uranium, l'invitation du FLI vise à réfléchir au moyen d'introduire un système de vérification de ces systèmes algorithmiques. Il s'agit de s'interroger sur les modes de gouvernance relatives à l'IA générative pour prendre le temps de co-construire un régime de vérification des laboratoires d'IA et de leurs systèmes. Plusieurs instances internationales comme l'OCDE, la Commission européenne, le Conseil de l'Europe ou l'UNESCO ont développé des recommandations et projets de réglementation spécifiques pour une IA de confiance qui méritent d'être soulignés. Nous allons développer les deux initiatives qui présentent une valeur juridique contraignante.

### Le projet de règlement de l'IA de la Commission européenne (AI Act)

La Commission européenne a présenté en avril 2021 à Bruxelles un projet de règlement sur l'intelligence artificielle (IA), appelé *AI Act*, dans le cadre d'une politique de réglementation du numérique visant à limiter les externalités négatives d'une approche par le marché [2]. Le Parlement européen votera la dernière version de ce texte le 27 avril 2023. Une question majeure dans les négociations est la manière de traiter les IA à usage général (GPAI), de grands modèles de langage qui peuvent être adaptés à diverses tâches. L'*AI Act* définit un système GPAI comme un système destiné par le fournisseur à exécuter des fonctions généralement applicables telles que la reconnaissance d'images et de la parole, la génération d'audio et de vidéo, la détection de formes, la réponse à des questions, la traduction, etc. Cette définition s'applique tant aux logiciels propriétaires qu'aux logiciels libres. Ces systèmes peuvent être utilisés dans plusieurs contextes et être intégrés dans plusieurs autres systèmes d'IA. Selon EURACTIV, il est désormais proposé d'inclure les GPAI dans la catégorie des systèmes à risque élevé et de les enregistrer dans la base de données de l'UE. Les ensembles de données qui alimentent ces grands modèles linguistiques devraient être soumis à des mesures de gouvernance des données appropriées. Tout au long de leur cycle de vie, les GPAI devront faire l'objet d'audits externes testant leur performance, leur prévisibilité, leur interprétabilité, leur corrigibilité, leur sécurité et leur cybersécurité, conformément aux exigences les plus strictes de l'*AI Act*. La responsabilité des acteurs serait également prise en compte tout au long de la chaîne de valeur de l'IA. Un nouvel article prévoit également d'empêcher les fournisseurs d'imposer unilatéralement des clauses contractuelles abusives aux PME. La liste des tâches de l'Office de l'IA de la Commission européenne a été élargie pour inclure des conseils sur la manière dont le règlement sur l'IA s'appliquerait aux chaînes de valeur de l'IA et sur les implications connexes en termes de responsabilité. L'urgence du Parlement européen à adresser la question de l'IA générative s'explique par le fort engouement pour ces systèmes ces dernières années. La lettre ouverte du FLI [11] qui demande une pause de 6 mois pour développer des protocoles de sûreté pour le design d'IA avancées, auditées et supervisées par des experts indépendants extérieurs, ainsi que des garanties appropriées, a également pu renforcer cette situation d'urgence. Si cette lettre ouverte a suscité une vive polémique [13], et si les actions proposées ont fait l'objet de critiques [9], elle présente le mérite d'accélérer la réflexion au niveau réglementaire sur le sujet. L'enjeu réside dans l'accroissement de la transparence et de la responsabilité.

### Le projet de Convention sur l'IA du Conseil de l'Europe

A Strasbourg, le projet international du Conseil de l'Europe vise à développer un standard juridique contraignant fondé sur les droits de l'homme, la démocratie et l'Etat de droit [8].

## Quelques défis de la prolifération des systèmes GPAI dans le domaine civil

La prolifération des systèmes GPAI présente des défis multiples : biais, désinformation, influence des processus électoraux, automatisation croissante des tâches et dévalorisation de la valeur travail, augmentation de la cybercriminalité et des cyber-menaces, enjeux énergétiques, autodétermination informationnelle, rareté des compétences, prolifération et vitesse de disruption des applications dérivées de ChatGPT comme notamment BabyAGI et auto-GPT, faible compréhension des décideurs

ou déni de réalité, etc. Ces enjeux nécessitent une gouvernance globale des risques, comme dans le domaine nucléaire ou éventuellement dans le domaine de l'aviation. La prochaine partie se concentrera sur deux défis : la propriété intellectuelle et la manipulation des comportements.

### Un défi pour la propriété intellectuelle

Les systèmes GPAI constituent un défi pour la propriété intellectuelle. A titre d'illustration, citons la plainte adressée par Getty Images contre Stability AI. Stability AI aurait copié plus de 12 millions de photographies des sites de Getty Images avec les légendes et les métadonnées et est aujourd'hui en concurrence directe avec Getty Images en commercialisant stable diffusion et son interface DreamStudio. Le tribunal a reconnu la violation du droit d'auteur et l'existence d'œuvres dérivées sans autorisation de la part de Getty Images. Des exceptions au droit d'auteur, prévues par la directive sur le droit d'auteur [1] à l'article 3, sont relatives à la fouille de textes et de données, définie à l'article 2.2, à des fins de recherche scientifique. Cette exception n'est cependant pas applicable, si l'utilisation des œuvres protégées n'a pas été expressément réservée par les titulaires de droits (article 4). Se pose dès lors la question de savoir dans quelle mesure une exception serait pertinente pour l'utilisation de données dans les modèles d'IA.

### Un risque de manipulation des comportements

L'IA générative peut être utilisée pour la génération de contenus (agents conversationnels, *deep Fake*, *fake news*) et de revenus publicitaires en manipulant les comportements [5]. A mesure que l'IA générative se répand dans des mains de plus en plus nombreuses, il devient de plus en plus difficile de déterminer qui manipule qui et selon quels calculs. Contrairement à l'Internet, l'IA générative ne se concentre pas uniquement sur l'information. Elle facilite certes la diffusion de l'information de manière exponentielle et remet en question la notion de vérité, mais a également le potentiel d'influencer les comportements et les décisions via des agents conversationnels. C'est probablement pour cela que Joe Biden a indiqué, lors de sa rencontre du Conseil consultatif sur la science et la technologie le 4 avril 2023, que l'IA générative pourrait être dangereuse [14].

Cette dangerosité concerne tant la prolifération des systèmes GPAI dans le domaine civil que dans le domaine militaire. Or, ni l'*AI Act* [2], ni le projet du conseil de l'Europe [8] ne s'appliquent au domaine militaire.

## Quelques défis d'une prolifération des systèmes GPAI dans le domaine militaire

Dans le domaine militaire, les armes dotées d'une autonomie sans intervention humaine soulèvent des préoccupations légales concernant notamment le régime de responsabilité applicable en cas d'incident. Elles soulèvent également des questions stratégiques complexes : la dépendance des États ou des entreprises à des systèmes d'intelligence artificielle a en effet le potentiel d'accroître la vulnérabilité de ces acteurs. Cela a été évoqué précédemment concernant l'accès aux capacités de fabrication des puces et semi-conducteurs.

Dans le même temps, la rapidité et l'efficacité des systèmes d'IA ont le potentiel d'augmenter l'efficience des attaques et d'accroître le marché de la surveillance. Des normes strictes sont nécessaires pour réduire les risques de prolifération des systèmes d'IA en particulier dans le domaine militaire et de la sécurité. Les technologies telles que la vision par ordinateur assistée par des algorithmes d'apprentissage automatique, la prise de décision auto-

nome alimentée par l'IA, les capteurs avancés et de plus en plus digitaux, les capacités informatiques miniaturisées à haute puissance, ainsi que les réseaux à haut débit seront au cœur de cette révolution [4]. Cet écosystème accélère la vitesse de la guerre et le risque d'escalade [3]. L'automatisation de la prise de décision et la simultanéité des actions en essaimage via l'IA ont pour effet la minimisation de la prise de décision humaine dans la grande majorité des processus traditionnellement nécessaires pour faire la guerre. Ces processus sont en cours d'évolution avec l'essor de l'IA générative [35]. Ils suscitent de nouvelles questions juridiques et éthiques en lien notamment avec l'application du droit de la guerre [36] et l'évolution de la tradition de la guerre juste (jus ad bellum et le jus in bello).

## Comparaison entre la prolifération des systèmes GPAI et la prolifération nucléaire

La prolifération des systèmes GPAI a le potentiel de nuire à la sécurité internationale de la même manière que la prolifération nucléaire. La possession d'armes nucléaires et la possession de systèmes GPAI créent en effet tous deux des risques pour la sécurité. Selon chatGPT, « la prolifération de l'IA pourrait également entraîner une surveillance de masse et la suppression des droits et libertés fondamentales ». Une course aux armements dans le domaine de l'IA favoriserait ainsi la concurrence pour le développement de technologies de pointe et amplifierait le risque de cyberguerres et d'attaques. Le rythme d'évolution de la puissance de calcul et la dissémination de l'usage de l'IA dans tous les secteurs d'activités obligent à repenser la doctrine stratégique. Est-ce qu'une agression « virtuelleé » justifie une force « cinétique » en réponse – dans quelle mesure et selon quelles équations d'équivalence ? Une nouvelle approche de la théorie de la dissuasion et de la doctrine stratégique devrait être élaborée [7].

Pour réglementer l'usage de ces technologies, il est nécessaire de réfléchir aux modes de gouvernance appropriés dans un monde multipolaire, dans le respect des cultures de chacun. Il est proposé de s'interroger sur l'opportunité d'accords de non-prolifération pour l'utilisation de l'IA dans le domaine militaire et de favoriser la coopération internationale sur ces sujets. Les États-Unis, l'Union européenne, le Japon et le Canada sont des États pionniers dans ce domaine et ont une grande importance dans les organisations internationales. La Chine pourrait également jouer un rôle clé dans le domaine des négociations sur une approche normative de l'IA sur le plan international dans le domaine militaire. Du fait de son expertise diplomatique et de la présence de nombreuses organisations internationales sur son territoire, la Suisse pourrait également faciliter les négociations d'un Traité dans le domaine de la non-prolifération de l'IA, sous l'égide des Nations-Unies. La motivation principale serait d'élaborer des pistes de réflexion pour prévenir la prolifération de l'IA générative dans le domaine militaire.

## Historique et enseignements de la réglementation sur la non-prolifération nucléaire

La fin de la guerre froide a réduit la menace d'une guerre nucléaire entre les superpuissances nucléaires, mais la prolifération technologique a augmenté la possibilité d'acquérir une capacité d'armement nucléaire, même pour les acteurs non étatiques. Le Traité de

35. Vice, Palantir Demos AI to Fight Wars But Says It Will Be Totally Ethical Don't Worry About It, April 26, 2023.
36. Palantir Blog, AI, Automation, and the Ethics of Modern Warfare, April 7, 2023.

Non-Prolifération Nucléaire (TNP), signé en 1968, a pour but d'empêcher la dissémination d'armes nucléaires. Il permet aux États non dotés d'armes nucléaires de recevoir de l'aide pour l'utilisation pacifique de la technologie nucléaire. Cependant, le TNP a eu du mal à s'imposer comme une norme internationale contraignante, et certains États, tels que l'Iran et la Corée du Nord, ont continué de développer des programmes nucléaires malgré les garanties du TNP.

Selon Kissinger [7], la prolifération des armes nucléaires risque de conduire à la formation de systèmes d'alliances comparables à ceux qui ont mené à la Première Guerre mondiale, avec une portée et une puissance destructrice bien plus importantes. Dans un monde nucléaire multipolaire, l'accès aux armes nucléaires est perçu comme un avantage stratégique, mais cela amplifie les risques de confrontation et de détournement, comme en atteste le développement des armes nucléaires en Iran et en Corée du Nord.

Il importe de relever que le TNP n'offre pas de mécanisme international défini pour faire respecter le traité en cas de violation des termes du traité ou de répudiation du traité par les États membres.

Le TNP a été suivi du Traité d'interdiction des armes nucléaires en 2017. Celui-ci interdit la possession et l'utilisation d'armes nucléaires, mais n'a pas été signé par les États nucléaires. Le traité *New Start*, prolongé par les États-Unis et la Russie en février 2021, limite les arsenaux nucléaires des deux pays, mais la modernisation des arsenaux et les programmes nucléaires de la Corée du Nord suscitent des inquiétudes quant à une nouvelle course aux armements nucléaires. Ainsi, les objectifs de désarmement et de non-prolifération n'ont pas été atteints. Si le TNP n'a pas eu les effets escomptés, la question se pose de savoir comment prévenir la prolifération de l'IA générative et dans quelle mesure une comparaison avec le droit international nucléaire serait pertinente pour le contrôle de la dissémination de ces systèmes ? Bien que la question reste ouverte, il pourrait être envisagé d'élaborer un traité de non-prolifération de l'IA pour empêcher la prolifération de l'IA à des fins autres que pacifiques. Cette perspective est partagée par Gary Marcus et Anka Reuel [37].

Quel pourrait être le champ d'application de ce traité ? Ce traité s'appliquerait à tout système informatique ou algorithmique pouvant accomplir des tâches normalement requises de l'intelligence humaine, telles que la résolution de problèmes, l'apprentissage, la prise de décisions, la compréhension du langage naturel et la reconnaissance d'images.

Les États signataires s'engageraient à ne pas développer, acquérir, stocker, déployer, utiliser ou transmettre intentionnellement des systèmes d'IA à usage préjudiciable ou destructeur. Les États signataires seraient tenus de garantir la transparence de leurs activités liées à l'IA, de favoriser la coopération internationale dans ce domaine et de mettre en place des normes éthiques et juridiques pour son utilisation à des fins pacifiques. Des mesures telles que des contrôles sur la production, l'utilisation, la vente ou le transfert d'IA, l'interdiction de l'exportation de l'IA à des pays non signataires, un système de vérification et de surveillance, ainsi qu'un programme de formation et de sensibilisation pour les décideurs, les développeurs et les utilisateurs d'IA seraient nécessaires pour assurer la conformité avec les dispositions du traité. Une organisation compétente devrait être créée pour faciliter la mise en œuvre du traité, qui serait ouverte à la signature de tous les États.

37. The Economist, The world needs an international agency for artificial intelligence, say two AI experts, April 18, 2023.

## Conclusion

En matière civile, la réglementation de l'IA est en cours de développement dans le cadre des institutions européennes et du Conseil de l'Europe. Des discussions sont en cours pour inclure les systèmes GPAI dans ce cadre normatif. Malgré le caractère extraterritorial de la réglementation européenne, et le nombre élevé d'États membres du Conseil de l'Europe, la question de l'encadrement normatif global de ces systèmes dans le domaine civil et militaire reste ouverte. Ce papier a exploré brièvement dans quelle mesure le droit international nucléaire pourrait servir de référence à l'élaboration d'un régime de vérification et de contrôle des systèmes GPAI, sous l'égide des Nations-Unies. La surveillance d'une autorité compétente indépendante disposant d'un mandat international pourrait être examinée de manière approfondie. La prolifération des armes nucléaires est devenue un problème majeur pour l'ordre international contemporain. Pour prévenir la prolifération des systèmes GPAI, une concertation internationale sous l'égide des Nations Unies serait nécessaire. Elle favoriserait un usage pacifique, uniquement défensif de l'IA, ainsi que le contrôle des armements sur la scène internationale. La légitimité démocratique du processus de réglementation est essentielle, de même que la mise en cause effective de la responsabilité des organisations pilotant le processus industriel. Cela inclut tous les acteurs. Selon Kissinger [7], plusieurs questions se posent : Que cherchons-nous à prévenir ? Que cherchons-nous à atteindre ? Quelle est la nature des valeurs que nous cherchons à promouvoir ? Répondre à ces questions nécessite la construction d'imaginaires pour un vivre ensemble commun, dans le respect des différences culturelles de chacun.

## Références

[1] Directive (UE) 2019/790 du Parlement européen et du Conseil sur le droit d'auteur et les droits voisins dans le marché unique numérique, 17 avril 2019.

[2] Proposition de règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (dit « AI Act »), COM/2021/206 final, 21 avril 2021. Voir la version actuelle de l'*AI Act* : Orientation générale du Conseil de l'Union européenne modifiant la proposition d'*AI Act*, document ST 15698 2022 INIT, 6 décembre 2022.

[3] J. R. Allen and D. M. West. It is time to negotiate global treaties on artificial intelligence. *Brookings*, March 2021.

[4] B. Buchnan. The AI Triad and What It Means for National Security Strategy. *Center for Security and Emerging Technology*, August 2020.

[5] M. Coeckelbergh, M. De Ketelaere, N. Smuha, P. Dewitte, and Y. Poullet. Open Letter: We are not ready for manipulative AI – urgent need for action. *KU Leuven - AI Summer School*, March 2023.

[6] N. Grant and K. Weise. In A.I. Race, Microsoft and Google Choose Speed Over Caution. *The New York Times*, April 2023.

[7] H. Kissinger. *World order*. Penguin Books, 2015.

[8] D. Leslie, Ch. Burr, M. Aitken, J. Cowls, M. Katell, and M. Briggs. Artificial intelligence, human rights, democracy, and the rule of law: a primer. 2021.

[9] S. Luccioni. The Call to Halt 'Dangerous' AI Research Ignores a Simple Truth. *Wired*, April 2023.

[10] Y. Meneceur and S. Sontag-Koenig. Réglementations européennes de l'intelligence artificielle : la tentation du pragma-

tisme. *Les Temps Électriques*, décembre 2022.

[11] Future of Life Institute. Pause Giant AI Experiments: An Open Letter, March 2023.

[12] M. Rochefort. Le Japon veut augmenter ses investissements de 82 % dans les semi-conducteurs. *Siècle Digital*, avril 2023.

[13] A. Seydtaghia. Un moratoire sur l'intelligence artificielle ? « Non, il est beaucoup trop tard ». *Le Temps*, mars 2023.

[14] B. Terrasson. L'IA dangereuse ? Pour Joe Biden « Cela reste à voir. Elle pourrait l'être ». *Siècle Digital*, avril 2023.

## Aspects normatifs de l'IA dans la Smart City : Optimisation, régulation et enjeux de gouvernance

Par **Laurence VANIN**
*RETINES / Docteur Es Lettres en philosophie politique et épistémologie*
*Université de Nice Côte d'Azur*
*Chaire Smart City : Philosophie et Éthique*
*Directeur du Comité d'Éthique de INR (Institut du Numérique Responsable)*
*Expert de l'Institut Europ IA*
*Expert du Smart Deal*
*Expert SAFE – Safe Cluster*
*Expert MWF – (Monaco)*
*RC – Lieutenant-Colonel de la Gendarmerie Nationale*
*laurence.vanin@univ-cotedazur.fr*
*http://www.laurencevanin.fr/*

### Introduction

La *Smart City* qui s'étend aujourd'hui aussi au *Smart* Territoire désigne la stratégie d'une ville ou d'un territoire dont les infrastructures sont gérées par des Intelligences Artificielles afin de faciliter les usages en matière de mobilité, d'économie d'énergie, de prévention des risques et de gestion des politiques territoriales. Nous constatons que la population urbaine est passée durant le XX$^{e}$ siècle de 220 millions à 2,8 milliards d'habitants et en 2050 plus de 70 pour cent de la population mondiale vivra dans des métropoles ou mégapoles. Mais pour autant, il ne faut pas déduire trop rapidement que les hommes déserteront totalement les campagnes pour les villes. La crise COVID a bien mis en évidence que la ruralité est aussi importante et qu'elle est plus que jamais reliée aux villes. Des citoyens résident sur tout le territoire et il est légitime qu'ils puissent bénéficier des mêmes services que les citoyens des villes.

Néanmoins, les chiffres « parlent » et impactent les décisions qui se calent sur les contraintes chiffrées, norment les choix. Ces chiffres qui étayent les effets d'annonce désignent ainsi un ensemble de signes qui suscite le questionnement sous-tendent parfois un discours alarmiste afin de faire changer les comportements. Ils peuvent sembler influant. De toute évidence, il importe de considérer les territoires de manière systémique et de concevoir différentes échelles de compréhension et de régulations des besoins liés aux activités des citoyens. Si autrefois, il fallait consulter son

compteur de consommation, prendre du temps pour réaliser de nombreuses démarches administratives, attendre ses factures de consommations d'énergie pour vérifier sa consommation, etc. Aujourd'hui les Intelligences Artificielles assistent les humains dans leurs démarches et des objets d'un genre nouveau ont fait leur apparition pour les y aider : les IoT. En quelques clics, il est possible d'accéder à ses consommations à partir d'un smartphone. C'est pourquoi les nouvelles politiques des *Smart* Territoires utilisent les technologies afin de parfaire leur efficacité. L'IA permet une gestion des consommations, comme des flux qui simplifient la tâche des techniciens, des administrateurs et des usagers.

Aujourd'hui, il importe de réfléchir sur les motivations qui animent les décideurs et acteurs des changements en matière de villes et territoires « Intelligents » car souvent ils justifient les besoins des habitants par rapport à des statistiques, des probabilités, des algorithmes prédictifs qui norment la création de *Smart City* ou territoires techno-centrés qui apparaissent alors comme des villes solutions et des territoires résilients. C'est pourquoi il importe de s'interroger sur ce qu'est une *Smart City* afin de comprendre comment l'IA est un véritable outil d'aide à la décision pour les villes désireuses de s'émanciper de leurs contraintes et en quoi les villages peuvent également prétendre à être *smart*. Néanmoins, si le territoire connecté demeure un espace de collecte des données, il importe aussi de comprendre comment elles sont traitées, ce qu'elles signifient. En quoi peuvent-elles être porteuses d'une signification « algorithmique » incompréhensible (ou inaccessible aux usagers qui ne sont pas familiers du langage informatique) et s'imposer comme des contraintes davantage que comme des opportunités. Il s'agit alors de comprendre en quoi l'aspect normatif de la donnée peut susciter des craintes. Mais qu'en est-il vraiment ?

## Des cités et des territoires smart : une optimisation au service des citoyens ?

Les chiffres justifient l'efficacité technique, qui elle-même sert le moyen d'accéder à une fin. Néanmoins, la réussite de la technique s'obtient au prix d'une économie de moyens pour un maximum d'efficacité. Certes, la technique modifie les conditions de vie, mais aussi impacte les façons de penser et les usages associés. D'autant qu'avec la Révolution industrielle[38], les performances techniques et la confluence entre l'Internet et l'IA ont multiplié les possibilités d'améliorer la qualité de vie, la santé en proposant un autre modèle économique. Ce constat mathématique [2] sur la concentration ou pas d'individus sur un périmètre restreint met en évidence d'autres facteurs – que ceux évoqués précédemment – auxquels le *smart* territoire va notamment devoir faire face : la mobilité, la sécurité, la concentration des besoins, etc.

La question de la mobilité a bien rapidement pris une avance sur les autres sujets, car elle touche à la nécessité de répondre à deux exigences : se déplacer sans difficulté et écologiquement. Les municipalités sont soucieuses de fluidifier le trafic urbain et périurbain, mais aussi de préserver la santé des usagers. La ville devient le lieu de l'entre-expression des réseaux de transports (bus, trams, métros, etc.) et des réseaux piétons (zones piétonnes et trottoirs interactifs).

Dans de nombreuses villes intelligentes sont

38. « De tous les défis multiples et fascinants auxquels nous sommes confrontés aujourd'hui, le plus important est de comprendre et orienter la nouvelle révolution technologique, qui n'implique rien de moins qu'une transformation de l'humanité. Nous sommes à l'aube d'une révolution qui bouleverse déjà notre manière de vivre, de travailler et de faire société. Ce phénomène inédit par son envergure et sa complexité constitue ce que je considère comme la Quatrième Révolution Industrielle » [12].

utilisés des capteurs embarqués afin de surveiller certains modèles de trafic routier et de zones d'embouteillages courantes afin d'améliorer les conditions de circulation. Les données collectées peuvent simplement mettre en évidence des dysfonctionnements et les solutions peuvent ainsi être trouvées afin de répondre au cas par cas aux réalités de terrain. Pour les espaces ruraux, il est essentiel de ne pas démunir les villages des services de transport nécessaires au maintien des populations dans les campagnes (bus, chemins de fer). Les technologies intelligentes permettent ainsi de repérer et de pourvoir à l'éventuelle détérioration des équipements comme les feux de signalisation et les panneaux lumineux pour les piétons ou véhicules. Elles quantifient aussi l'effet du trafic sur l'environnement, la Nature.

La concentration élevée du nombre d'individus au cœur des villes pose aussi des problèmes d'insécurité. Les villes intègrent donc la demande sécuritaire des habitants, c'est pourquoi la vidéo surveillance, par exemple, est en plein essor. La cybersécurité connaît également une réelle évolution avec l'intégration de technologies de reconnaissance faciale à ces caméras, pour l'identification d'individus suspects ou dangereux. En plus de la reconnaissance faciale, les caméras de vidéo surveillance dernier cri sont également équipées de détecteurs de mouvements et de fumée, et pourvues d'alarmes incendie. Ces villes envisagent de s'équiper de « lanceurs » d'alarme afin de réduire le temps d'intervention des services de secours. L'installation de boutons d'alarme fixes à travers la ville facilitera aux forces de l'ordre le repérage précis de leur lieu d'intervention et permettra de diriger la circulation à l'aide des technologies intelligentes pour arriver plus rapidement à destination. Dans les campagnes, la gendarmerie propose un service de proximité afin de répondre aux besoins des habitants. La concentration d'individus est moins importante pourtant toutes les zones du territoire sont sensibles, dès lors que les services informatiques doivent garantir les libertés individuelles et la gestion de données de qualité. Bien d'autres idées sont mises en application en vue d'améliorer la sécurité publique.

Le *Smart* Territoire projette d'offrir des conditions de vie optimales réunies dans un habitat totalement adaptable conciliant efficacité énergétique et confort thermique en vue du bien-être des résidents. Pour cela, il est important que chacun puisse bénéficier des réseaux de connections fiables et performants. La *Smart City* envisage de changer d'allure avec des architectures mêlant le verre et le vert afin de manifester le désir de transparence et le besoin d'accorder une place prépondérante aux végétaux. Si les espaces sont reconsidérés avec une verticalité favorable à la construction de nombreux logements, elle ouvre aussi à la création de jardins suspendus. Un visuel qui semble alors incarner la quête d'un certain retour à la nature par l'entremise d'une [39] et par une hyper végétalisation des façades, des espaces communs.

Par ailleurs, l'attention portée aux données collectées a pour effet de générer un réaménagement du territoire par la transformation des services et un changement de politique de la ville et des villages connectés. Elle tend à faire évoluer le concept de *Smart City* au territoire dans une vision tentaculaire. La *Smart City* s'appuie donc sur les technologies et les

39. « Pourquoi une pelouse ? "Parce que c'est joli, une pelouse", pourrait répondre le couple. Mais pourquoi les trouve-t-il jolies ? C'est l'aboutissement d'une longue histoire. Les chasseurs-cueilleurs de l'âge de pierre ne cultivaient pas l'herbe à l'entrée de leurs cavernes. Aucune verte prairie n'accueillait les visiteurs de l'Acropole athénienne, du Capitole à Rome, du Temple de Jérusalem ou encore de la Cité interdite à Pékin. L'idée de faire pousser une pelouse à l'entrée des demeures privées et des bâtiments publics est née dans les châteaux des aristocrates français et anglais à la fin du Moyen-Âge » [7].

capacités de l'industrie digitale afin de se réapproprier l'espace urbain et de favoriser chez les citoyens une prise de conscience susceptible de modifier leurs comportements pour qu'ils deviennent coproducteurs des services dont ils ont besoin. En conséquence, les données [4] entrent en jeux et initient l'évolution des raisonnements qui intègrent de plus en plus les menaces à venir afin d'anticiper sur les « risques » ou les catastrophes. Le pouvoir des données recueillies s'impose, car elles ne mentent pas. Elles disent une autre réalité et rendent visibles des phénomènes qui ne peuvent laisser indifférents (pic de pollution, nombre d'hospitalisations qui s'ensuivent, qualité de l'air, etc.). Dès lors, les ingénieurs et acteurs du territoire se saisissent de ces éléments afin d'influer les politiques de la ville et de formuler [8] un projet de *Smart City* écoresponsable susceptible de devenir pour tous *une évidence*. Dès lors, les campagnes ne peuvent se réduire à de vastes espaces susceptibles d'approvisionner les villes en énergie, au détriment des paysages. De la même manière, il est difficile pour les citoyens de voir les services se réduire progressivement avec la réduction des services « humains prodigués par des humains » là où les télécabines médicales sont installées en remplacement d'un médecin de famille. Les territoires ne peuvent donc souffrir de traitements inégaux sans engendrer la colère des citoyens qui se désespèrent de l'arrivée de technologies normées par les résultats, les indicateurs, et non par le qualitatif.

## Du nombre au numérique : un territoire connecté au service des hommes ou une entrave à leur liberté ?

Avec l'Internet des Objets, la ville tout entière devient connectée afin de créer des services inédits qui octroient aux divers acteurs de la ville – collectivités et opérateurs – la possibilité d'être plus efficients dans leurs activités. Grâce à son fort taux de numérisation, la ville devient de plus en plus attractive pour les individus. Elle propose un meilleur niveau de services (transports, loisirs...), une plus grande sécurité sans oublier un cadre de vie plaisant. La ville est aujourd'hui, et certainement plus encore demain, en concurrence avec ses paires du monde entier. Elle se doit donc d'être performante. Ainsi, le numérique est-il devenu un des vecteurs essentiels de cette compétitivité.

Les informaticiens se sont associés aux ingénieurs et urbanistes afin de traduire numériquement les données et de proposer une manière inédite de traiter l'information. L'objectif *Smart City* a accentué la vitesse d'innovation dans les entreprises et transformé la vision des dirigeants. Le concept a engendré la disruption [40].

La transformation numérique a permis aux entreprises d'intégrer toutes les technologies digitales disponibles au sein de leurs activités. Ces procédés ont pour vocation d'améliorer la performance et la production des entreprises et favorisent leur croissance. Elles souhaitent répondre aux besoins des clients et surtout délivrer les bonnes prestations au bon moment. Pour y parvenir, elles désirent connaître leurs

40. Jean-Marie Dru, Président non exécutif du groupe de communication américain TBWA est l'auteur du concept de Disruption. La version française de son livre « New » rappelle sa définition initiale, qui n'est pas une théorie pour décrire l'existant, mais bien plutôt « une méthodologie dynamique tournée vers la création » [3]. C'est l'idée qui permet de remettre en question les « Conventions » habituellement pratiquées sur un marché, pour accoucher d'une « Vision », créatrice de produits et de services radicalement innovants. « L'innovation disruptive est une innovation de rupture, par opposition à l'innovation incrémentale, qui se contente d'optimiser l'existant », précise Jean-Marie Dru. Pour Bernard Stiegler à l'inverse « la disruption est ce qui va plus vite que toute volonté, individuelle aussi bien que collective, des consommateurs aux "dirigeants", politiques aussi bien qu'économiques » [13].

clients et futurs clients et cherchent à être visionnaires dans ce monde en mutation. Savoir ce qui fonctionnera demain est à leur portée, connaître les nouveaux modes de consommation également et c'est au travers des actions humaines et des informations que les individus cèdent jour après jour que cela est faisable. En conséquence, la transformation numérique des entreprises accompagne les changements sociétaux. C'est pourquoi elles anticipent afin de répondre aux besoins futurs. Un des gros enjeux actuels pour les acteurs de la *Smart city* consiste en la capacité à faire parler les données que les entreprises ont accumulées durant de nombreuses années.

En réalité, la transition numérique a facilité l'interface entre les systèmes, la mutualisation des services et a montré que les algorithmes sont une autre manière de dire la ville, en anticipant sur les émotions et en automatisant les probabilités [41]. Désormais, les villes vont disposer d'un mobilier urbain connecté. C'est pourquoi les usagers prennent conscience de la valeur des données, des datas, d'autant que tous n'ont pas accès au langage informatique. Chacun s'interroge donc sur les liens qui s'établissent entre l'utilisateur et le collecteur de données. « Dans l'univers illimité des flux, l'État érige des murs visibles, terrestres et sous-marins qu'il veut le moins visible possible » [10] Et l'enjeu éthique repose sur le cadre juridique qui entoure la donnée et qui nécessite d'interroger les *process* (stockage, utilisation, exploitation), mais aussi la sécurisation et la gouvernance. L'arrivée de l'IA également laisse perplexe puisqu'un ensemble de tâches qui revenait initialement à l'humain lui échappe et est réalisé par des machines. « Au XX$^{e}$ siècle, le libéralisme aura beaucoup plus de mal à se vendre. Alors que les masses perdent leur importance économique, l'argument moral seul suffira-t-il à protéger les droits de l'homme et les libertés ? Les élites et les gouvernements continueront-ils à apprécier la valeur de chaque être humain sans que cela rapporte le moindre dividende économique ? Dans le passé, il y avait quantité de choses que seuls les humains pouvaient faire. Désormais, robots et ordinateurs rattrapent leur retard et pourraient bientôt surpasser les hommes dans la plupart des tâches. Certes, les ordinateurs fonctionnent tout autrement que les hommes, et il est peu probable que des ordinateurs s'humanisent dans un futur proche. En particulier, il est peu probable que des ordinateurs soient sur le point d'acquérir une conscience et se mettent à éprouver émotions et sensations. L'intelligence informatique a accompli d'immenses progrès au fil du demi-siècle écoulé, mais la conscience des ordinateurs n'a absolument pas progressé »[13].

Qu'il vive à la campagne ou au cœur d'une ville, l'homme se connecte. Pour son travail, se tenir informé ou encore communiquer avec ses contemporains, les humains s'emparent des réseaux sociaux parce qu'Internet leur simplifie l'existence. Ils peuvent passer des commandes, jouer, communiquer, vivre des expériences immersives, voir des concerts en ligne, etc. De la même manière, l'IA paraît pouvoir anticiper sur les aspirations des hommes et les rendre si prévisibles qu'ils se trouvent dépourvus d'une part de leur libre arbitre. L'IA a inventé l'homme « prédictible ». « Transformés en fournisseurs de data, ceux-ci (les individus et

41. « Au cours des dernières décennies, les biologistes ont acquis la conviction que l'homme qui appuie sur les boutons et boit le thé est lui aussi un algorithme. Sans doute un algorithme beaucoup plus compliqué que le distributeur automatique, mais un algorithme quand même. Les humains sont des algorithmes qui produisent non pas des gobelets de thé, mais des copies d'eux-mêmes (comme un distributeur automatique qui, si vous pressiez la bonne combinaison de boutons, produirait un autre distributeur). Les algorithmes qui régissent les distributeurs opèrent via des engrenages et des circuits électriques. Ceux qui contrôlent les humains opèrent par les sensations, les émotions et les pensées » [7].

les groupes que les réseaux dits "sociaux" déforment et re-forment selon de nouveaux protocoles d'association) s'en trouvent désindividués par le fait même : leurs propres données, qui constituent aussi ce que l'on appelle (dans le langage de la phénoménologie husserlienne du temps) des rétentions, permettent de les déposséder de leurs propres propensions – c'est-à-dire de leurs propres désirs, attentes, volitions, volonté, etc. » [7].

La combinatoire de nombreuses données produit des résultats tendant à dissoudre les désirs des hommes et leur liberté dans un algorithme qui devient le référent au détriment de l'humain, qui initialement était défini comme être de liberté. La confiance qui est accordée à des magmas de code, « d'existences autonomes » entraîne un nouveau genre d'« auto accroissement de la technique », qui produit un effet singulier : une forme d'éloignement à l'égard des humains. En effet, l'algorithme utilise les données fournies par les hommes, mais l'IA calcule dans la froideur des automatismes informatiques et impose ce qui aura valeur de décision. Entre gain de temps et optimisation des services, l'IA facilite le quotidien des usagers, mais elle calcule froidement du fond de ses serveurs, souvent délocalisés et produit des « prédictions » qui influent le devenir des utilisateurs, privés de la spontanéité de leur choix.

## Smart Territoire, vérité et transparence des données : vers une IA éthique

Si la transition numérique sous-tend le concept de *Smart* territoire, le territoire connecté propose en un langage informatique « imperceptible » une formulation qui rend compte d'une vérité. Elle peut alors rendre manifeste une réalité que les algorithmes révèlent sans que chacun puisse « les voir » à l'œuvre dans le processus d'imitation de l'intelligence humaine [1, 6]. « Désormais, la charge dévolue au numérique ne consiste plus seulement à permettre le stockage, l'indexation et la manipulation aisés de corpus chiffrés, textuels, sonores ou iconiques en vue de diverses fins, mais à divulguer de façon automatisée la teneur de situations de tous ordres. Il s'érige comme une puissance aléthéique, une instance vouée à exposer l'aléthéia, la vérité, dans le sens défini par la philosophie grecque antique, entendu comme le dévoilement, la manifestation de la réalité des phénomènes au-delà de leurs apparences. Il se dresse comme un organe habilité à expertiser le réel de façon plus fiable que nous-mêmes autant qu'à nous révéler des dimensions jusque-là voilées à notre conscience » [11].

Ce *logos* technique donne à voir, dans son prolongement, une réalité, des résultats qui influent sur les agissements des hommes et empruntent au lexique des sciences cognitives. L'algorithme impacte donc le monde sensible et situe la vérité au-delà des apparences, dans ce qui ne change pas au rythme des fluctuations de la condition humaine. À cela s'ajoute les défis climatiques et la peur s'installe progressivement. Certes les hommes souhaitent changer leurs habitudes de consommation, mais le monde techno-centré les « emprisonne » aussi paradoxalement dans d'autres pratiques énergivores. Les chiffres, les algorithmes, les indicateurs sur fond de volonté d'optimisation imposent de nouvelles pratiques sociales. « En cela, il prend la forme d'une techné logos, une entité artéfactuelle douée du pouvoir de dire, toujours plus précisément et sans délai, l'état supposé exact des choses. Nous pourrions affirmer que nous entrons dans le stade achevé de la technologie, ne désignant plus un discours portant sur la technique, mais un terme qui prendrait acte de sa faculté à proférer du verbe, du logos, mais dans l'unique but de garantir le vrai. Ce pouvoir constitue la caractéristique première de ce qui est nommé "intelligence artificielle", déterminant à la suite toutes fonctions qui lui sont assignées » [11].

La technique trouve ainsi sa finalité dans la vérité d'un langage informatique et se révèle alors dans son abstraction puisque les objets ne contiennent rien. Tout est délocalisé dans des serveurs. L'IA aurait alors vocation à dire un *Logos* qui exprimerait une vérité qui demanderait à être dévoilée. En imitant les fonctions cognitives et en développant certaines des aptitudes habituellement réservées à l'humain, l'IA incarnerait ce dispositif qui dépasse les capacités humaines et forme une organisation systémique du traitement de l'information et facilite la coïncidence avec le réel. Plus rapide dans la réalisation des calculs, elle se veut parangon du système neuronal [42] et marque ainsi l'apparition de l'aspect anthropologique de la technique. D'autant qu'elle semble aussi gagner en autonomie grâce au *deep learning*. La machine peut alors traiter de nombreuses informations et se dispenser des hommes pour le faire. « D'abord, c'est un anthropomorphisme augmenté, extrême ou radical, cherchant certes à se modéliser sur nos capacités cognitives, mais les prenant comme des leviers afin d'élaborer des mécanismes qui, tout en s'inspirant de nos schémas cérébraux, sont voués à être plus rapides, efficaces et fiables que ceux qui nous constituent, tout en étant tendanciellement inaltérables. Ensuite, c'est un anthropomorphisme parcellaire ; il n'a pas vocation à embrasser la totalité de nos facultés cognitives et à traiter, comme nos esprits, une infinité de questions, étant seulement destiné, en l'état actuel des choses, à assurer des tâches spécifiques. Enfin, c'est un anthropomorphisme entreprenant qui ne se contente pas d'être doué de seules dispositions interprétatives, mais est envisagé comme une puissance capable d'engager, de façon automatisée, des actions en fonctions de conclusions arrêtées » [11].

Certes, ces anthropomorphismes expriment une « vérité » des hommes et consistent en l'imitation, en mode accéléré, de leurs facultés de raisonner. Cette accélération ou rapidité de l'IA à calculer et produire ses résultats imite les facultés cognitives et renforce sa crédibilité et son efficacité. Cependant, il nous faut pousser l'argument à son terme. Effectivement, nous pourrions également ajouter à ce triple devenir anthropomorphique de la technique un quatrième anthropomorphisme dit « métaphysique et normatif ». Tendant à être exponentiel, universel, nouménal, inaccessible car – métaphysique – en deçà de la physique, il aurait une fonction injonctive « normative », ce qui renforcerait son efficacité. La logique de l'algorithme tendrait alors à s'appliquer à l'ensemble des actions de la vie collective ou individuelle en soumettant et en infléchissant les décisions ou les êtres à sa puissance de calcul. L'homme ne peut pourtant s'aliéner à la sophistique informatique. Il se doit de la reconsidérer et de la ranger parmi les opinions qu'il incombe de soumettre à la méthode ou logique afin de l'éprouver dans le but de la confirmer ou l'infirmer. En conséquence, en confiant une partie de leurs tâches à la logique algorithme, les hommes se dessaisissent d'une part de leurs réflexions, mais se doivent malgré tout de conserver une part de leur capacité de décision et d'exercice

42. « Nous sommes donc capables de visualiser le cerveau vivant avec des modèles générés par ordinateur (et non pas directement !) sans intervention massive dans l'organe même. Dans le communiqué, cette avancée en médecine est liée à une promesse plus vaste, celle de rendre visible le fait de penser, une proposition illusoire, et, disons-le carrément, un non-sens. On peut effectivement rendre visible des processus cérébraux, mais pas l'acte de penser lui-même » [5].

43. « Thomas Anderson, alias Néo, n'aura existé que comme un programme informatique, et sa réalité ontologique aura été nulle. Sorti d'un rêve simulé pour suivre un lapin blanc dans son terrier, Thomas a cru choisir entre la pilule rouge de la réalité et la pilule bleue du simulacre sans savoir que son choix était déjà programmé. Sa lutte contre les machines se sera limitée à une série de contrôles internes de la Matrice effectuée par l'anomalie du système informatique. La paix que Néo a instaurée avec les machines est aussi fallacieuse que le rêve dont il a

de leur libre arbitre, au risque de se mettre en danger[43]. La responsabilité ne peut donc être transférée à la seule puissance de l'IA, l'humain doit donc garder au cœur des *Smart* Territoires et des processus mis en œuvre dans leur élaboration, sa capacité à raisonner et à finaliser les décisions afin d'en rester maître. Après réflexion, il ne peut s'aliéner à « la série des abstractions » en s'en remettant à la puissance métaphysique et normative de la technique. D'autant que cette puissance contribue aussi à sa souveraineté. En d'autres termes, la technique – souvent considérée comme neutre – n'a pas été soumise à la contrainte morale puisqu'elle avait vocation à se déployer sans limites pour servir quelques fins. C'est pourquoi elle ne peut dans le cadre du *Smart* Territoire se développer de manière totalement libre, sans être régulée, contrôlée. La puissance et l'autonomie de la technique ne peuvent, comme le redoute Ellul dans son ouvrage *Le système technicien*, justifier que la technique « soit juge de la morale ». La *Smart City* ne peut devenir le théâtre de la mise en œuvre des technosciences où « une proposition morale ne sera considérée comme valable pour ce temps que si elle peut entrer dans le système technique, si elle s'accorde avec lui ».

*In fine*, la morale ne peut être assujettie à la technique, comme en son projet Machiavel dans *Le prince* plaçait l'État au-dessus de la morale au point d'en conclure que la fin justifiant les moyens : « Le vice paraîtrait vertu. » La réalisation de la *Smart City* et des *Smart* Territoires questionne donc la problématique du devoir.

## Conclusion

Le concept de *Smart City* doit être déconstruit afin d'être remodelé sur la base d'autres arguments que ceux évoqués jusqu'ici. En quoi, outre les prouesses techniques qu'elle entend développer, le territoire du futur pourra-t-il faire coexister des Valeurs ? Plus qu'une rhétorique plaidant en faveur de son utilité, des principes peuvent-ils servir à modéliser le concept de *Smart* Territoire ? Il est essentiel d'intégrer des problématiques servant l'exemplarité, l'égalité entre tous que l'on réside dans un village ou dans une mégapole afin de permettre à chacun de bénéficier de nouveaux services, mais aussi d'user de son droit à la déconnexion pour vivre – par moment – en parfaite adéquation avec la Nature.

## Références

[1] J.-M. Besnier. *L'homme simplifié. Le syndrome de la touche étoile*. éd. Fayard, 2012.

[2] S. Caird, I. Hudson, and G. Kortuem. A tale of Evaluation and Reporting in UK Smart Cities. *The Open University, UK*, 2016.

[3] J.-M. Dru. *New : 15 approches disruptives de l'innovation*. Pearson, 2016.

[4] FNCCR. Guide pratique et notices juridiques à destination des collectivités territoriales et de leurs groupements. Les enjeux du big data territorial, 2016.

[5] M. Gabriel. *Pourquoi je ne suis pas mon cerveau*. éd. JC Lattès, p. 21, 2017.

[6] J.-G. Ganascia. *Le mythe de la singularité. Faut-il craindre l'intelligence artificielle ?* éd. Seuil, 2017.

[7] Y. N. Harari. *Homo Deus, une brève histoire de l'avenir*. éd. Albin Michel, 2017.

[8] R. Laugier. La ville de demain : intelligente, résiliente, frugale, post-carbone ou autre. In *synthèse documentaire, Paris,*

cru s'évader. Il mourra donc, à la fin de la trilogie, non pas comme un héros de la liberté, mais comme la fonction d'un programme parvenu à son terme » [9].

*CRDAL, Ministère de la transition écologique et solidaire et Ministère de la cohésion des territoires*, 2013.

[9] J.-F. Mattéi. *L'homme dévasté*. éd. Grasset, p. 145, 2015.

[10] O. Mongin. *La ville des flux. L'envers et l'endroit de la mondialisation urbaine*. éd. Fayard, p. 196, 2013.

[11] E. Sadin. *L'intelligence artificielle ou l'enjeu du siècle, anatomie d'un antihumanisme radicale*. éd. L'échappée, 2019.

[12] K. Schwab. *La quatrième révolution industrielle*. éd. Dunod, p. 11, 2017.

[13] B. Stiegler. *Dans la disruption. Comment ne pas devenir fou*. éd. LLL, 2016.

## Quelles normes pour l'IA en médecine ?

Par **Serge TISSERON**
*Psychiatre / Docteur en psychologie HDR*
*Président de l'Institut pour l'étude des relations homme-robots (IERHR)*
*Membre de l'Académie des technologies*
*Membre du Conseil national du numérique (CNNum)*
*Membre du Conseil scientifique du CRPMS (Université de Paris Cité, ED 450)*
*Co-responsable du DU de Cyberpsychologie (Université de Paris Cité)*
*serge.tisseron@gmail.com*

En médecine, l'IA présente de nombreux avantages à la fois pour l'aide au diagnostic et pour les thérapies [8]. Mais à vouloir la développer sans prendre en compte ses impacts sur le terrain, la médecine peut rapidement se déshumaniser. De plus, certains objets numériques peuvent développer ou aggraver la souffrance psychique, soit parce qu'ils ont été conçus sans autre souci que les profits de leur fabricant, soit parce que certains usagers les utilisent de façon excessive ou maladroite, que ce soit de leur propre initiative ou parce que cela leur est imposé. C'est pourquoi l'utilisation de l'IA pose des défis considérables en lien avec la santé mentale, tant au niveau individuel que collectif, qui imposent de fixer des normes qui en balisent les usages. Les domaines concernés sont notamment la protection des données des utilisateurs contre le vol et le piratage, l'influence de ces technologies sur les comportements des patients en dehors des séances, l'égalité d'accès à ses ressources, et les conditions d'une alliance thérapeutique entre cliniciens et utilisateurs de services dans un environnement numérique.

### La menace d'une médecine déshumanisée

Ce qu'il est convenu d'appeler la *médecine computationnelle* fait sans cesse de nouveaux progrès, qui sont également sensibles dans le domaine de la santé mentale. On désigne sous cette expression trois champs fortement interdépendants [4] : l'utilisation d'outils connectés pour recueillir des données numériques sur les patients dans leur vie quotidienne ; la création de grandes bases de données et leur analyse par apprentissage machine de façon à perfectionner la différenciation et la catégorisation des troubles ; et le développement de modèles mathématiques du fonctionnement et du dysfonctionnement afin de mieux comprendre les mécanismes à l'origine des symptômes observés en clinique.

En pratique, l'IA permet d'ores et déjà de

construire des plans de traitement individualisé pour chaque patient, intégrant les données qui sont disponibles sur lui, sans forcément prendre en compte la façon dont le patient se perçoit lui-même [3]. Le médecin soigne le double numérique du patient (son « jumeau ») et vérifie régulièrement par de nouveaux examens si celui-ci va mieux, sans se poser la question de savoir si l'original est désespéré, dépressif, voire suicidaire, sauf si un dosage biologique opportunément demandé lui révèle son état d'esprit, pour autant que cela soit possible. Pour beaucoup de patients habitués à voir un médecin, il peut en résulter le sentiment d'être abandonné à des machines [7]. Quant aux soignants, si ce système est d'une efficacité avérée en termes de rentabilité du temps médical, il semble en revanche accélérer l'épuisement de ceux qui se sont orientés vers cette profession par attrait pour la relation de soin [9], avec des risques de *burn out* accrus.

Dans un avenir proche, l'intelligence artificielle permettra à des robots d'intervenir auprès de patients sans aucune intervention humaine : pour réaliser le diagnostic, voire pour engager un traitement, y compris dans le domaine psychologique, comme le montrent déjà les ambitions de Facebook avec *Woebot*, son robot « coach en santé mentale » [2]. Un engagement qui ne relève pas d'un idéal philanthropique, loin de là. Rappelons en effet que cette entreprise s'enrichit de la capture des données personnelles de ses usagers, qu'elle utilise ou revend. Cela assure déjà *Woebot* de bénéficier de beaucoup d'informations pour poser les bonnes questions à ceux qui décident de l'utiliser : ce robot exploite tout ce que son utilisateur et ses proches ont mis à propos de lui sur le réseau social. Mais en plus, les confidences qui lui sont faites constituent autant de nouvelles données personnelles que Facebook pourra exploiter !

Enfin, une ambition associée à ces technologies est de ne pas seulement informer sur le présent d'un patient, mais aussi de prévoir son évolution [6]. Dans ces conditions, la stratégie du « prendre soin », centrale dans la pratique soignante, se trouve bouleversée. Alors que la médecine et la psychiatrie se sont construites avec comme objectif de soulager une souffrance perçue par le patient, les examens pratiqués sur lui et les études de cohortes peuvent orienter sur une autre voie que celle de ses aspirations personnelles, y compris dans le domaine de la psychologie : l'inviter à accepter des techniques que des algorithmes désignent comme pouvant améliorer son état à long terme, indépendamment de la perception qu'il a de lui-même et de sa santé [7]. Pour éviter cette dérive, il est urgent de créer des normes invitant à privilégier les machines qui favorisent l'intelligibilité des algorithmes et les liens entre soignants et patients, et plus largement entre usagers. C'est la condition pour développer une médecine de la personne globale et sociale, et pas seulement une médecine de l'individu réduit à ce que des machines peuvent mesurer de lui à un moment donné.

## Pour un virtuel sensible, protecteur et respectueux

L'enjeu majeur du développement des technologies numériques en matière de diagnostic et de soin, y compris médico-psychologique, réside dans la conscience que les soignants doivent garder de l'écart entre deux réalités : d'un côté, les informations que ces technologies leur donnent sur un individu qui n'est semblable à aucun autre, et les moyens d'agir sur ses dysfonctionnements ; et d'un autre côté, le sujet réel, pétri de son histoire, de ses attentes et de ses doutes.

### Une médecine de la personne et pas de l'individu

Aucun patient n'est semblable à autre du point de vue de sa génétique et de sa biologie intime, c'est pourquoi il justifie de soins spécifiques. Mais ce processus d'individuation de chaque patient doit être complété en lui donnant les moyens d'intégrer l'expérience personnelle qu'il fait de la maladie dans sa vie mentale pour qu'elle s'y transforme en savoir sur lui-même, lui permettant de gérer sa vie en y intégrant sa maladie. Or, tant qu'une expérience émotionnelle et cognitive n'est pas interprétée, critiquée, transformée, ce n'est pas un savoir : le savoir commence lorsque les informations sont classées, organisées, ordonnées, interprétées, critiquées, et surtout transformées par un individu ou un groupe pour être mises au service d'une dynamique psychique, individuelle ou collective. Le moteur de l'appropriation du savoir passe par la reformulation [1]. Il appartient donc au numérique de favoriser la possibilité d'une reformulation, par le patient, de ses expériences vécues, qu'il faut lui permettre d'anticiper en lui apportant toutes les informations nécessaires avant l'examen ou au début de la thérapie.

### Intelligibilité des algorithmes

De plus en plus de diagnostics médicaux qui ont des implications sur nos vies dépendent du résultat de systèmes algorithmiques. Ces algorithmes mettent en œuvre, le plus souvent de manière opaque, des critères de priorité, de préférence, de classement qui ne sont généralement pas connus des personnes concernées. Aujourd'hui, le médecin et le malade sont dans l'ignorance de la hiérarchie des informations que les machines leur donnent et des raisons pour lesquelles elles les leur donnent. Le grand public a découvert ce problème avec les réponses que ChatGPT faisait à leurs questions, mais le problème est beaucoup plus général, et particulièrement préoccupant lorsque des décisions de santé doivent être prises, que ce soit pour un individu ou un groupe. On parle parfois de la nécessité de rendre les algorithmes transparents, mais, pour les utilisateurs, le fonctionnement d'un algorithme a peu d'intérêt. L'intelligibilité est plus importante que la transparence. Autrement dit, il s'agit de transmettre aux usagers toutes les informations utiles pour qu'ils puissent en interpréter les résultats. Et pour cela, il faudrait contraindre les concepteurs des algorithmes d'aide à la décision de produire, outre des résultats attendus, des éléments d'explication sur ce qui a prévalu dans les choix qui ont été faits, à la fois pour les soignants et pour les patients.

### Des machines au service de la création des liens

Les robots seront de plus en plus capables de répondre aux attentes de communication simples, telles que partager des conseils de cuisine, jouer à un jeu ou coacher des exercices physiques. Mais un robot peut également informer son utilisateur sur les ressources humaines de proximité. Par exemple, l'existence d'un club de quartier ou de personnes elles aussi isolées avec lesquelles l'usager pourrait entrer en contact pour s'adonner à son activité préférée. Les fabricants de robots doivent se voir imposer de concevoir des robots qui favorisent les relations entre les humains, tout comme ils sont obligés de créer des robots qui ne mettent pas en danger la santé physique de leurs utilisateurs. Dans les deux cas, il y va de la santé, mentale d'un côté et physique de l'autre. Autrement dit, le modèle du robot de compagnie doit être le robot « humanisant » qui contribue aux rencontres entre humains plutôt que le robot humanoïde capable de se substituer à un humain de compagnie [5].

## En conclusion

L'IA appliquée à la médecine doit être mise au service d'un accroissement de la capacité des patients à comprendre leur pathologie, à l'accepter et à envisager sereinement les thérapeutiques proposées. Elle pourrait être l'occasion de créer une forme de co-immersion par laquelle le thérapeute accompagnerait le patient dans sa découverte d'aspects inconnus de lui-même, afin qu'il puisse mieux se connaître et faire évoluer ses capacités et ses compétences selon ses choix. Ce droit est également une condition d'une société libre permettant à chacun de se réaliser conformément à ses souhaits. C'est pourquoi le développement de l'intelligence artificielle doit s'accompagner de normes destinées à en baliser les usages. À défaut, ce que l'on appelle « la médecine de la personne » restera en réalité une médecine de l'individu, menacée par la technicisation, l'hyperspécialisation et, finalement, la perte de sens, tant pour les soignants que pour les patients, entraînant un risque de déshumanisation et de maltraitance, sur un chemin pourtant pavé de bonnes intentions. Les patients sont des personnes et les soignants aussi. Ils ont un nom. L'intelligence artificielle, elle, n'en a pas. Une médecine anonyme est ce qu'il y a de pire.

## Références

[1] D. Brixhe and A. Specogna. Actes de reformulation et progression du savoir. *Pratiques : linguistique, littérature, didactique*, 103-104 :9–27, 1999.

[2] K. K. Fitzpatrick, A. Darcy, and M. Vierhile. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot). A randomized controllled trial. *JMIR Mental Health*, 4(2), 2017.

[3] K.J. Friston, K.E. Stephan, R. Montague, and R.J. Dolan. Computational psychiatry : the brain as a phantastic organ. *Lancet Psychiatry*, 1(2) :148–158, 2014.

[4] C. Gauld, G. Dumas, E. Fakra, J. Mattout, and Micoulaud-Franchi J.-A. Les trois cultures de la psychiatrie computationnelle. *Annales médico-psychologiques*, 179(1) :63–71, 2021.

[5] S. Tisseron. *Le jour où mon robot m'aimera. Vers l'empathie artificielle*. Albin Michel, Paris, 2015.

[6] S. Tisseron. *L'emprise insidieuse des machines parlantes*. Les Liens qui libèrent, Paris, 2020.

[7] S. Tisseron. L'intelligence artificielle (IA), promesses et inquiétudes : une médecine anonyme est ce qu'il y a de pire. *Médecine et philosophie*, 4(2) :25–30, 2020.

[8] S. Tisseron and F. Tordo (Dir.). *Pratiquer les cyberpsychothérapies : jeux video, réalité virtuelle et robots*. Dunod, Paris, 2022.

[9] J.-M. Triffaux, S. Tisseron, and J. Nasello. Decline of empathy among medical students. Dehumanization or useful coping process ? *Encéphale*, 45(1) :3–8, 2021.

AfIA
Association française
pour l'Intelligence Artificielle

# IA, neurosciences et technologies : tension entre liberté citoyenne et liberté de la recherche scientifique. Premiers résultats d'une démarche de science participative

Par

**Éric FOURNERET**
*ETHICS EA 7446 / ETH+*
*Université Catholique de Lille*
*eric.fourneret@univ-catholille.fr*

**David DOAT**
*david.doat@univ-catholille.fr*

**Nathanaël LAURENT**
*Sciences, Philosophies et Sociétés / ESPIN*
*Université de Namur*
*nathanaël.laurent@unamur.be*

**Yves POULLET**
*Namur Digital Institute / ESPIN*
*Université de Namur*
*yves.poullet@unamur.be*

**Valérie TILLMAN**
*Sciences, Philosophies et Sociétés / ESPIN*
*Université de Namur*
*valerie.tilman@unamur.be*

**Mathieu GUILLERMIN**
*CONFLUENCE Sciences et Humanités*
*Université Catholique de Lyon*
*mguillermin@univ-catholyon.fr*

## Introduction : neurosciences, IA et identité humaine

L'hypothèse centrale du projet de science participative « *New Humanism at the time of Neuroscience and Artificial Intelligence* » (NH-NAI), dont nous présentons dans cet article la démarche et un premier résultat, postule que les progrès de ces dernières années en intelligence artificielle et dans les neurotechnologies réinterrogent en profondeur nos conceptions de ce que signifie pour chacun et collectivement « être humain ».

En effet, si les avancées en IA et dans les technosciences du cerveau portent l'espoir d'importantes avancées dans le domaine de la santé, dans la compensation des formes de handicap (*i.e.* « *Brain Computeur Interfaces* », BCI) et dans le traitement de certains troubles neurologiques (diagnostic et approches thérapeutiques), elles annoncent aussi des applications (*i.e. monitoring*, traçage, profilage ou surveillance) qui pourraient s'insérer au plus intime de la vie des sujets et dans leur vie sociale, ou impacter en profondeur les moyens de l'éducation (tel que l'usage de ChatGPT en formation) et les pratiques démocratiques. De ce

fait, elles soulèvent de nombreuses questions quant à notre identité humaine et son avenir dans un monde où le rôle des technologies numériques dans l'évolution des sociétés ne fait plus de doute.

Tout citoyen étant concerné par ces nouvelles questions, le projet NHNAI entend offrir un cadre de réflexion collective et de délibération qui permettra une élaboration de pistes et propositions de réponses communes. Il accorde à la pluralité des points de vue une considération significative, appréhendée par divers moyens : organisation d'entretiens individuels, mise en place d'ateliers-citoyens en présentiel et tenue de débats publics en ligne via une plateforme numérique (« Cartodébat ») où chaque citoyen peut entrer en dialogue et participer à l'échange des arguments [44]. L'intérêt du projet NHNAI est ainsi de susciter et permettre un dialogue au sein d'un échantillon de citoyens porteurs d'une diversité de savoirs, de compétences et d'engagements en société, pour dégager des lignes de conduites enrichies de la diversité des apports de toutes les parties prenantes de la discussion publique.

Sur cette base, l'objectif final du projet est de parvenir, au terme de son programme en 2024, à formuler des recommandations à destination des acteurs politiques et des institutions, dans l'espoir que ce type de technosciences n'impacte pas la société sans que cette dernière n'en ait collectivement pensé quelques enjeux éthiques et sociaux. Ces recommandations ne seront pas issues d'un référendum ou d'un sondage, ni d'une « moyenne » des opinions collectées. L'enjeu du projet est de penser une approche éclairant et informant d'autant mieux le politique dans ses décisions que ce dernier sera, outre l'avis des plus experts sur un sujet (les scientifiques et représentants en sciences humaines et sociales), accompagné par des recommandations élaborées et délibérées collectivement par des citoyens faisant part de leurs propres préoccupations.

Dans le cadre de cet article, nous nous limitons à situer le projet NHNAI par rapport à la littérature en sciences participatives. Nous présentons ensuite un premier résultat des échanges entre participants du projet, qui souligne leurs préoccupations en matière d'IA, de respect de leur liberté et de préservation des conditions les meilleures pour la réalisation de la recherche. Nous proposons enfin une première interprétation de ces résultats à la lumière des concepts d'aliénation et de privauté mentale.

## « Wicked problems » et sciences participatives

À l'instar des sciences participatives et des conférences de consensus, le projet NHNAI part du postulat que les applications de l'intelligence artificielle et des neurotechnologies soulèvent un questionnement quant à l'identité humaine et aux conditions de sa compréhension et de son devenir, qui ne peut être éclairé sans tenir compte des dimensions collective et interdisciplinaire.

Cette exigence existe en raison de la spécificité du problème traité, qui relève de la catégorie du *wicked problem* (en français : problème difficile ou noueux) dans la littérature anglo-saxonne. Un problème est difficile lorsque « les faits sont incertains, les valeurs débattues, les enjeux élevés et les décisions urgentes » [12], mais aussi parce que toutes les parties concernées, scientifiques, experts et société publique, sont touchées par ce problème (ils en font partie), mais ne sont pas *a priori* d'accord sur ses enjeux, ses causes et ses conséquences, voire sur le type de stratégie nécessaire pour les résoudre [13]. Pour ces raisons, la réso-

44. Ces contenus sont coconstruits par les membres de NHNAI à partir des premiers entretiens individuels et des débats citoyens qui se sont tenus à Lille et Namur entre mai et septembre 2022.

lution d'un *wicked problem* requiert le franchissement d'une prétendue frontière classique entre science et société, avec la participation de toutes les parties prenantes à la production de connaissances partagées [11].

La recherche conduite dans le cadre de NHNAI s'inscrit dans la continuité de cette littérature, en considérant que les avancées technoscientifiques en intelligence artificielle et dans les neurosciences, s'ils manifestent indéniablement des aides tout à fait remarquables, posent aussi des problèmes qui affectent l'ensemble de la société, soulevant des questions pratiques d'ordre éthique qui ne peuvent être résolues par les seuls moyens d'une expertise pluridisciplinaire entre les sciences et les Sciences Humaines et Sociales. Ces questions touchent à la justice sociale, au bien commun, au respect des valeurs et principes d'une société démocratique, au respect des personnes et de leurs libertés, aux conditions de stabilité du contrat social, aux finalités collectivement souhaitables d'un projet de société. C'est dans ce contexte que la notion d'expertise collective, pluri-professionnelle, pluridisciplinaire et pluriculturelle (les acteurs du projet NHNAI proviennent d'Europe, des Amériques, de l'Asie et de l'Afrique), apparaît centrale dans la démarche de recherche entamée par les acteurs du projet NHNAI. Reconnaissant aux individus non-experts la capacité de penser le « bien-commun » et la justice, le projet NHNAI considère leurs arguments comme les prémisses d'un raisonnement public respectueux du pluralisme dans nos sociétés et limitant le conflit d'intérêts au niveau de la réflexion éthique qui, sinon, risquerait d'être cantonnée dans un « entre-soi institutionnel ».

Cela étant, et comme dans toute démarche de réflexion collective, des limites existent, ce qu'atteste la littérature sur le questionnement relatif à la crédibilité de la narration de citoyens (quel que soit leur niveau de formation, leur statut social ou symbolique, leurs engagements professionnels et disciplinaires) sur des sujets aussi pointus et complexes. Comme le souligne Jürg Steiner [15], le danger est de favoriser une logique de contestation – où le dialogue serait visé dans un second moment – plutôt que délibérative. Cet écueil bien connu a le mérite de rappeler malgré tout qu'une réflexion éthique privée de certaines expressions de la vie morale, parce qu'elles ne seraient pas exclusivement rationnelles, laisserait de côté une partie importante de nos expériences d'êtres humains, comme celles qui sont liées à la dimension affective. En effet, toute conception du « Bien » soutenue par des arguments rationnels n'en est pas moins un attachement à ce qu'on désire personnellement ou collectivement voir se réaliser.

Dès lors, puisque chacune et chacun sont déjà, peu ou prou, immergés dans le monde numérique, et confrontés à ce que les neurosciences comprennent de l'être humain, le projet NHNAI inscrit ses analyses dans une approche soucieuse des préoccupations scientifiques et citoyennes, en créant un espace et une méthode grâce auxquels toutes les voix sont entendues et prises en compte par les partenaires du projet. Dans ce cadre, les justifications de points de vue qui n'empruntent pas toujours les chemins classiques et attendus de la rationalité ne souffrent pas nécessairement de manque de légitimité. En effet, une bonne délibération, rappelle Jane Mansbridge [9], ne peut pas exclure toutes les logiques de raisonnements sans, en même temps, manquer à l'exigence éthique d'écouter toutes les expressions de la vie morale. D'une part, il s'agit de reconnaître que le sentiment est une composante intégrale de la raison et qu'il est impossible au jugement de s'y soustraire [8]. D'autre part, toute personne possède non seulement une idée du bien commun et de la justice, mais aussi la capacité d'en justifier sa conception.

## L'inquiétude pour l'aliénation

Les expressions citoyennes dans les premiers entretiens et ateliers participatifs NHNAI ont fait apparaître de nombreux questionnements. Dans les limites du présent article, nous nous limiterons aux enjeux soulevés autour de la notion de liberté.

Sur des aspects démocratiques, des participants reconnaissent que les outils numériques, tels que les réseaux sociaux, ont permis positivement d'ouvrir des espaces dans lesquels toutes les voix peuvent se faire entendre, cette pluralité des voix soutenant ainsi le jugement critique : « On a plusieurs réseaux [numériques] et donc du coup le fait de confronter les réseaux, ça permet à chacun de mettre à l'épreuve l'information qu'il reçoit. » Mais cette reconnaissance du caractère positif des innovations numériques s'accompagne aussi d'un questionnement critique.

Un certain nombre d'échanges concerne des thématiques largement appréhendées dans la littérature (telle que la protection des données personnelles). Leur analyse laisse aussi apparaître, tant sur le thème de la démocratie que sur celui de la santé et de l'éducation, la crainte d'une forme d'« aliénation » dont on peut rendre compte de la façon suivante : les technologies numériques et les systèmes d'IA permettent aux humains de faire plus de choses et plus rapidement, mais au lieu de leur libérer du temps ou d'accroître leur dimension humaine, ce temps d'exécution gagné est perçu comme l'opportunité d'ajouter de nouvelles actions à accomplir, souvent au bénéfice de tierces parties, ou de mettre l'humain de côté. Autrement dit, l'aliénation capture l'idée d'un état de dépossession de soi au profit de quelque chose d'autre qui peut être une personne au sens physique et juridique, un système artificiel ou une personne morale (entreprise, association, État, etc.)

Prenons le cas de la numérisation du partage d'informations et de connaissances. Des participants perçoivent ce processus comme un démultiplicateur des actions humaines, principalement en termes de communication. Si la machine permet de faire plus de choses, plus rapidement, voire plus efficacement, s'exprime aussi la crainte d'une « suractivité » au détriment d'un temps libéré pour soi et pour l'autre : « J'ai l'impression [. . . ] [qu'] il va y avoir une espèce de multiplicité de pleins de choses, de plus en plus d'évènements et comme on pourra être partout à la fois, il [n'] y aura plus de limite à la suractivité humaine qui est déjà beaucoup trop intense. » Par ailleurs, ils s'interrogent sur le sens du recours aux technologies numériques quand celles-ci présagent un remplacement ou un dépassement de l'humain. Par exemple, les neurotechnologies équipées avec des IA, afin de modifier le cerveau en apportant des compensations à des fonctions cognitives perdues (*i.e.* la mémoire, si cela venait à exister réellement), ne finiraient-elles pas par poser une question d'identité ? « J'ai l'impression que du coup [. . . ] c'est essayer [. . . ] de repousser les limites d'un être humain pour [le] faire devenir autre chose qui [. . . ] n'est plus humain. »

Sur d'autres aspects, l'usage de l'IA et des technologies numériques apparaît comme un renforcement possible des interactions humaines. Il en est ainsi de la relation de soin quand des participants envisagent que la délégation de gestes techniques à des robots équipés d'IA puissent la ré-humaniser (*i.e.* « robots-chirurgiens »), réservant la présence des soignants au profit de l'écoute de leurs patients. Les participants les conçoivent plus performants que leurs homologues humains, ces derniers étant « désagréables parce qu'ils sont pressés, et donc le côté humain est mis de côté. [. . . ] [Avec les robots-chirurgiens], la personne qui s'occupera du patient sera plus humaine, car plus dans le contact, car elle ne passera pas seulement cinq minutes avec le patient. »

Mais « arrivera-t-on encore à former des super-chirurgiens [humains], si on laisse les compétences aux machines ? » Ne serait-ce pas perdre en capacités d'initiative ?

Pour toutes ces raisons, certains participants estiment qu'un cadre réglementaire est nécessaire, sans être pour autant une réponse suffisante autour de ces technologies. En effet, ce type de démarche vise à normer les actions humaines pour empêcher des dérives. Mais elle pourrait aussi avoir des effets négatifs, par exemple, sur la recherche scientifique. C'est un avis largement partagé parmi les participants : « Je pense que la régulation du développement technologique n'est pas forcément une mauvaise idée. Mais cela a tendance à brider les chercheurs dans leurs recherches, ce n'est pas bon pour eux. Il faut les encourager à chercher au maximum et non empêcher la recherche. » Comment protéger la liberté dans la société civile sans altérer les conditions de la réalisation de la recherche ?

Mais la crainte pour la ré-humanisation de la relation de soin finit finalement par l'emporter en imaginant l'éloignement des acteurs du soin par le renforcement d'une santé excessivement technicisée, qui ne laisserait plus de place à l'empathie : « Je crains que le monde de la santé ne devienne un monde déshumanisé, où le *soin* sera géré par le diktat d'algorithmes et de machines, à coup de données statistiques et de prédictions, où la préoccupation majeure sera de plus en plus de faire du business, donc de réduire les coûts, donc de faire des compressions de personnel où, de ce fait, les médecins et autres soignants manqueront cruellement, où la santé deviendra une donnée *objectivable*, gérable à distance par visioconférence ou autre procédé numérique, sans aucune place pour le ressenti du sujet. » La relation de soin, fondée en grande partie sur la clinique, deviendrait-elle une relation centrée exclusivement sur des aspects techniques, et le consentement d'un patient un « clic » des CGU (« Conditions Générales d'Utilisation »), comme ceux opérés déjà sur internet ?

Une telle tension se retrouve aussi dans le domaine de l'éducation. D'un côté, les progrès des connaissances dans le fonctionnement cérébral et ceux des outils informatiques aident à améliorer les programmes éducatifs et les stratégies pédagogiques : « L'intelligence artificielle et les neurosciences pourraient apporter [...] un outil pour aider plus individuellement les gens en leur offrant des techniques plus adaptées à un suivi personnalisé. » D'un autre côté, les outils numériques « envahissent » tout le secteur éducatif, laissant moins de choix en faveur d'une éducation diversifiée et modérée quant à leurs usages – que l'on pense, par exemple, au recours croissant des étudiants à ChatGPT3, aux usages et gains de temps remarquables qu'ils laissent entrevoir, mais aussi à ses effets d'objectivation, d'anonymisation et de standardisation des connaissances. La place de l'humain dans des processus de formation de plus en plus numérisés est donc questionnée : « Si on donnait le savoir par une IA, est-ce que l'enseignant aurait encore une place ? » Le soutien numérique à l'éducation ne risque-t-il pas également de s'accompagner d'un appauvrissement par le « formatage de l'humain [devenant davantage] un pion dans la société de consommation » ? L'élève sera-t-il encore maître de son parcours scolaire ? Peut-on parler d'autonomie dans ce contexte, laissent entendre certains participants NHNAI désireux d'interroger le rôle de l'éducation dans sa capacité à renforcer l'apprentissage de la liberté dans son rapport au savoir. Un tel questionnement n'est pas sans rappeler une réflexion philosophique entamée au moins depuis Marx, Bergson et Arendt [10, 3, 2], sur les rapports entre le travail et la machine.

Ainsi, les premiers entretiens et ateliers du projet NHNAI laissent apercevoir à travers les

questionnements qui s'en dégagent des positions contradictoires touchant à la liberté, tant en matière de démocratie, que de santé et d'éducation. Une certaine autonomisation de dispositifs artificiels équipés d'IA semble favoriser une délégation de plus en plus conséquente d'actions auparavant réservées à l'humain, avec de nombreux bénéfices mais au risque d'une réduction de sa liberté. En tenant compte de ces tensions, nous pouvons envisager que l'un des enjeux serait de penser des solutions (en matière de *design* responsable, *ethics by design*, par exemple) permettant de tenir ensemble les services que de tels dispositifs peuvent rendre aux humains, tout en préservant les capacités d'initiatives de ces derniers.

## Conclusion : liberté et privauté mentale

Si les universitaires, les gouvernements et les entreprises ont conscience que le développement des IA, des neurotechnologies et l'accroissement des connaissances dans le fonctionnement cérébral soulèvent de profondes questions éthiques et sociales, les premiers entretiens et ateliers que nous avons organisés montrent que cette conscience est aussi largement partagée par le grand public, révélant un besoin aux contours encore difficile à dessiner de l'intégration de la science et de la société comme une condition nécessaire du développement des technologies numériques, des IA et des connaissances. La science est certes conditionnée par des décisions de financement institutionnelles et en cela, l'activité scientifique est profondément sociale. Mais le grand public attend aussi qu'elle soit normée par des besoins et des valeurs sociaux plus larges que les seules sphères des avancées technologiques et des gains économiques perçus sous une forme de retour sur investissement dans le monde de l'industrie. C'est typiquement ce qu'expriment certains participants NHNAI dans leurs entretiens en soulevant leurs préoccupations concernant leur liberté, souvent comprise comme la liberté de l'esprit.

Certes, ils ignorent ce qui existe déjà institutionnellement, à l'image du programme « Recherche Responsable et Innovation » (RRI, « *Responsible Research and Innovation* », 2013) [1], et dont le principal objectif est d'assurer l'équilibre entre les conditions nécessaires à la réalisation de la recherche scientifique et la nécessité de protéger les libertés des citoyens. Mais l'accent mis sur la notion d'« aliénation » attire notre attention sur une préoccupation forte exprimées par les acteurs sociétaux rencontrés, que l'on peut traduire par un autre concept massivement utilisé dans la littérature de langue anglaise, à savoir la « privauté mentale » (« *mental privacy* ») [7, 14, 4]. Cette dernière désigne au moins quatre caractéristiques de l'esprit humain : i) la liberté de penser et/ou de vouloir (liberté cognitive) ; ii) le contrôle que le citoyen exerce sur ses données personnelles issues de son d'esprit, telles que les idées politiques, les croyances, l'expérience intérieure des émotions (intimité mentale) ; iii) le fonctionnement du cerveau et ses parties (intégrité mentale) ; iv) l'identité personnelle (continuité psychologique) [6, 5].

Cette notion de privauté mentale ne pourrait-elle pas être élevée au registre d'une norme (éthique mais aussi juridique) ? D'une façon très générale, bien des normes éthiques s'incarnent positivement dans la sphère juridique au sens kelsien du terme, à travers les notions de licite et d'illicite. La norme se présente alors sous la forme d'une contrainte externe qui s'exerce sur les individus par la menace de la sanction. Si la privauté mentale était alors reconnue comme une norme, ne pourrait-elle pas constituer une piste de réponse possible aux préoccupations citoyennes, quant à la question de savoir ce qu'il est moralement souhaitable ou non de faire à l'esprit d'un individu au moyen de technologies du cerveau et de dis-

positifs artificiels équipés d'IA ? La constitution d'une telle norme ne permettrait-elle pas en effet de poser les bases d'une préservation de l'intégrité de l'esprit, tout en offrant les conditions d'une recherche sereine ?

Une analogie avec le principe éthique et juridique de l'inviolabilité du corps humain, consacré dans le droit français (article 16-1 du Code civil), peut aider à mieux saisir la portée de cette proposition soumise au débat. Avec un tel principe, l'intégrité du corps humain se trouve juridiquement protégée. D'une part, ce qui peut être fait au corps peut être sanctionné si on lui a porté illicitement atteinte. D'autre part, le principe d'inviolabilité du corps ne permet pas qu'il ait le statut d'un objet patrimonial : il ne relève pas du droit de la propriété et ne peut être un bien appropriable (comme le sont les choses). Autrement dit, le droit considère que toute atteinte au corps est en même temps une atteinte à la personne : se situant entre l'être et l'avoir, le corps n'a pas en effet les qualités d'un objet dont nous pourrions nous séparer.

Au regard des paroles des participants NHNAI, nous pouvons nous demander si l'inviolabilité du corps humain, élevée au registre d'une norme juridique, ne peut pas constituer une analogie féconde pour imaginer les bases d'un droit protecteur de « l'inaliénabilité » de la privauté mentale de la personne humaine (*i.e.* son caractère non appropriable, qui ne peut faire l'objet d'une aliénation). Ne devons-nous pas concevoir la privauté mentale comme fut pensée celle du corps humain, dans un contexte contemporain de mutations technologiques rapides, aux effets incertains, où l'esprit demande qu'on lui prête attention ? Corps et privauté mentale, inviolabilité et inaliénabilité, ne sont-ils pas les deux faces d'une même pièce, d'un nouvel humanisme à l'ère des neurosciences et du développement des IA ?

Les analyses de certains entretiens et débats citoyens qui ont eu lieu au sein du programme NHNAI suggèrent que la norme de privauté mentale pourrait constituer, à condition d'en élaborer les critères opératoires et ses limites scientifiques, un rempart contre certaines formes d'aliénations liées aux mésusages des technologies numériques et former, à l'instar des bornes d'un fleuve, une balise éclairante pour soutenir et orienter le développement des IA et des neurotechnologies dans un cadre respectueux de l'intégrité humaine. Ce ne sera pas seulement aux scientifiques et aux politiques d'en décider, la société civile doit être entendue dans un échange démocratique autour de valeurs communes.

## Références

[1] *European Commission, Directorate-General for Communication, Directorate-General for Research and Innovation, Responsible research and innovation (RRI), science and technology : report, Publications Office,* https://data.europa.eu/doi/10.2777/4572, 2013.

[2] H. Arendt. *Condition de l'homme moderne*. Calmann-Lévy, éd. Pocket, Paris, p. 199–200, 1994.

[3] H. Bergson. *Les deux sources de la morale et de la religion*. PUF, Paris, p. 327, 1932.

[4] M. Enserink and G. Chin. The end of privacy. *Science*, 347(6221) :490–491, 2015.

[5] M. Ienca and R. Andorno. Towards new human rights in the age of neuroscience and neurotechnology. *Life Sciences, Society and Policy*, 13(5) :2–27, 2017.

[6] M. Ienca and P. Haselayer. Hacking the brain : brain-computer interfacing technology and the ethics of neurosecu-

rity. *Ethics and Information Technologies*, 18(2) :117–129, 2016.

[7] A. J. Kolber. Pain detection and the privacy of subjective experience brain imaging and the law. *American Journal of Law & Medicine*, 33(2–3) :433–457, 2007.

[8] S. R. Krause. *Civil Passions. Moral Sentiment and Democratic Deliberation*. Princeton University Press, 2008.

[9] J. Mansbridge, J. Bohman, S. Chambers, D. Estlund, A. Føllesdal, A. Fung, C. Lafont, B. Manin, and J. L. Martí. The place of self-interest and the role of power in deliberative democracy. *The Journal of Political Philosophy*, 18 :64–100, 2010.

[10] K. Marx. *Le Capital, 1867-1879, t. 1, 4e section, chapitre XV, 3, p. 379 (trad. J. Roy)*. Édition du Progrès, 1976.

[11] M. Mormina. Knowledge, expertise and science advice during covid-19 : In search of epistemic justice for the 'wicked' problems of post-normal times. *Social Epistemology*, 36(6) :671–685, 2022.

[12] C. Pohl, B. Truffer, and G. Hirsch-Hadorn. *Addressing wicked problems through transdisciplinary research*. R. Frodeman, J. Thompson Klein, and R. C. S. Pacheco (Eds), Oxford University Press, 4th edition, The Oxford handbook of interdisciplinarity : 319–331, 9780198733522.013.26, 2017.

[13] F. Popa, M. Guillermin, and T. Dedeurwaerdere. A pragmatist approach to transdisciplinarity in sustainability research : From complex systems theory to reflexive science. *Futures*, 65 :45–56, 2015.

[14] F. X. Shen. Neuroscience, mental privacy, and the law. *Harvard Journal of Law & Public Policy*, 36(2) :653–714, 2013.

[15] J. Steiner. Raison et émotion dans la délibération. *Archives de philosophie*, 72(2) :259–274, 2011.

# L'utilisation d'outils de machine learning à des fins de sécurité publique : une interdiction de principe en droit européen ?

Par

**Yves POULLET**
*CRIDS / NADi*
*Université de Namur*
*yves.poullet@unamur.be*
*www.unamur.be*

**Michael LOGNOUL**
*michael.lognoul@unamur.be*

## Introduction

L'intelligence artificielle (IA) constitue un outil majeur d'investigations, de prévention et de lutte contre la criminalité et le terrorisme. Si la liste des applications possibles de l'IA en matière policière ou de renseignement est infinie, notons que la technologie de l'apprentissage machine requiert cependant des mégadonnées et que la collecte, le stockage et l'exploitation de ces données peuvent être d'autant plus difficiles qu'elles sont à l'origine recueillies par des opérateurs privés. En outre, les textes en ma-

tière de protection des données, en particulier la directive 2016/680 (dite directive « Police-justice ») [2], fixent à cette exploitation et aux investigations policières nombre de contraintes.

Dans ce contexte, deux décisions majeures de la Cour de justice de l'Union européenne retiendront notre attention. La première, rendue en date du 6 octobre 2020, concerne les obligations de rétention des données de communication par les opérateurs de communication électroniques et leur utilisation par les autorités policières [4]. Ces obligations peuvent être imposées auxdits opérateurs par les États Membres de l'Union européenne (UE), en vertu de la directive 2002/48 (dite directive « *e-Privacy* », en cours de révision) [1]. Dans cet arrêt, la Cour soulève les risques accrus liés aux utilisations potentielles des technologies de l'IA, en tout cas celles utilisant les technologies de l'apprentissage machine, pour justifier un encadrement plus strict de l'étendue de ces obligations. La seconde décision date du 21 juin 2022 [6]. Elle concerne la transmission obligatoire, aux autorités publiques compétentes par les compagnies aériennes, de données relatives à leurs passagers, sur base de la directive 2016/681 (dite directive « PNR ») [3]. L'utilisation de logiciels d'IA par les services de police et de renseignements des États Membres, pour le traitement de telles informations, amène la Cour à interpréter de manière restrictive le texte européen, mais surtout à énoncer quelques principes en ce qui concerne cette utilisation. Dans les pages qui suivent, les points principaux de ces développements sont soulignés.

## La rétention des données et l'obligation de collaboration des opérateurs de communications électroniques

Dans cette première affaire, la Cour de justice était saisie par diverses associations de défense des libertés, notamment de la question suivante : « L'article 15, paragraphe I, de la directive [2002/58], combiné avec les articles 4, 7, 8, 11 et 52, paragraphe I, de la [Charte], doit-il être interprété en ce sens qu'il s'oppose à une réglementation nationale [. . . ] qui prévoit une obligation générale pour les opérateurs et fournisseurs de services de communications électroniques de conserver les données de trafic et de localisation au sens de la directive [2002/58], générées ou traitées par eux dans le cadre de la fourniture de ces services si cette réglementation a notamment pour objet de réaliser les obligations positives incombant à l'autorité en vertu des articles 4 et 7 de la Charte, consistant à prévoir un cadre légal qui permette une enquête pénale effective et une répression effective de l'abus sexuel des mineurs et qui permette effectivement d'identifier l'auteur du délit, même lorsqu'il est fait usage de moyens de communications électroniques ? » (§ 79).

Dans son raisonnement, la Cour consacre cette obligation positive de l'État, qui trouve un écho dans l'article 15, paragraphe 1, de la directive *e-Privacy* ; cette obligation trouve cependant ses limites dans l'application du principe de proportionnalité. L'obligation positive dont question permet certes aux États Membres d'introduire des exceptions à l'obligation de principe, énoncée à l'article 5, paragraphe 1, de cette directive, de garantir la confidentialité des données à caractère personnel ainsi qu'aux obligations de non-utilisation des données à des fins autres que de sécurité du réseau ou de facturation des services. Ces exceptions ne peuvent cependant valoir que si elles constituent une mesure prévue par la loi, nécessaire, appropriée et proportionnée, au sein d'une société démocratique, pour sauvegarder la sécurité nationale, la défense et la sécurité publique, ou assurer la prévention, la recherche, la détection et la poursuite d'infractions pénales ou d'utilisations non autorisées du système de communications électroniques.

Notre propos n'entend pas analyser l'en-

semble des règles déduites par la Cour de justice en ce qui concerne les limites du droit des États Membres soit à exiger des opérateurs de communication l'accès non généralisé, mais à certaines données de communication (par exemple, provenant d'une zone géographique particulière ou d'un groupe de personnes), soit à intercepter des communications. Il se concentre sur celles relatives à la conservation exigée des opérateurs de communication électroniques, des métadonnées de communication. Par métadonnées de communication, on entend les données de trafic et de géolocalisation (type de communication, émetteur, destinataire, localisation de ces acteurs, durée de la communication, volume des données) permettant d'identifier les communications sans atteindre à leur contenu. Cette obligation de conservation autorise alors les services de polices ou de renseignements à procéder par des techniques de *data mining* et de profilage à détecter les auteurs d'infraction, potentiels, suspectés ou réels. Ces opérations sont susceptibles de révéler des informations sur un nombre important d'aspects de la vie privée des personnes concernées, y compris des informations sensibles, telles que l'orientation sexuelle, les opinions politiques, les convictions religieuses, philosophiques, sociétales ou autres ainsi que l'état de santé. On sait que ces données méritent une protection particulière, suivant les textes européens. Par ailleurs, l'agrégation des métadonnées de communication peut aboutir à la constitution de profils très précis, incluant les habitudes de la vie quotidienne, les lieux de séjour permanents ou temporaires, les déplacements journaliers ou autres, les activités exercées, les relations sociales de ces personnes et les milieux sociaux fréquentés par celles-ci, et permet d'en inférer le contenu des communications.

Ce sont précisément les possibilités croissantes d'atteintes à la vie privée liées à l'utilisation de systèmes d'IA de plus en plus performants qui, selon la Cour, justifient des restrictions supplémentaires au droit des États Membres d'exiger une conservation généralisée des données de trafic et de géolocalisation, ce qui était l'objet du recours pris contre la réglementation de certains États Membres qui autorisaient cette demande de conservation. L'arrêt souligne les risques d'erreurs, de discrimination, et d'évolutivité non contrôlée liés à l'utilisation de tels systèmes (nous reviendrons sur ce point *infra*). La Cour relève en outre qu'une telle mesure concerne tous les citoyens et non uniquement ceux suspectés ou objets de mesure de surveillance et exige donc des restrictions supplémentaires. Aussi, les juges décident que de telles mesures doivent rester tout à fait exceptionnelles et ne s'adresser qu'à des mesures dites de sauvegarde de la sécurité nationale, à savoir la lutte contre le terrorisme. Ils excluent dès lors le recours à une obligation de conservation généralisée pour des objectifs de simple sécurité publique (par exemple, des manifestations violentes) ou de lutte contre la criminalité, y compris grave. Pour autant que la menace s'avère réelle et actuelle ou prévisible, et à la condition que la durée de cette conservation soit limitée au strict nécessaire, l'objectif de sauvegarde de la sécurité nationale face à une menace grave est seul susceptible de justifier des mesures comportant des ingérences dans les droits fondamentaux plus graves que celles que pourraient justifier ces autres objectifs. La Cour ajoute que le niveau de la menace, les techniques d'analyse automatisée et la durée de la mesure doivent faire l'objet d'un contrôle effectif « soit par une juridiction, soit par une entité administrative indépendante, dont la décision est dotée d'un effet contraignant, visant à vérifier l'existence d'une situation justifiant ladite mesure ainsi que le respect des conditions et des garanties devant être prévues » (§ 179). Les autorités de protection des don-

nées sont implicitement visées pour effectuer ce contrôle.

Face aux risques liés à l'utilisation des techniques d'IA, les juges énoncent quelques garanties supplémentaires qui doivent précisément faire l'objet de l'examen par cette autorité dont l'intervention est jugée nécessaire. Ainsi, « il convient de préciser que les modèles et les critères préétablis sur lesquels se fonde ce type de traitement de données doivent être, d'une part, spécifiques et fiables, permettant d'aboutir à des résultats identifiant des individus à l'égard desquels pourrait peser un soupçon raisonnable de participation à des infractions terroristes et, d'autre part, non discriminatoires » (§ 180). À cet égard, ils mettent en garde contre l'utilisation de modèles qui se fonderaient exclusivement sur des données sensibles comme l'origine raciale ou ethnique, les opinions politiques, les convictions religieuses, l'appartenance syndicale, l'état de santé ou la vie sexuelle d'une personne, sans prendre en compte l'analyse du comportement individuel de la personne. Le taux d'erreurs constatées à la suite de l'utilisation des systèmes d'intelligence artificielle exige que « tout résultat positif obtenu à la suite d'un traitement automatisé doit-être soumis à un réexamen individuel par des moyens non automatisés avant l'adoption d'une mesure individuelle produisant des effets préjudiciables à l'égard des personnes concernées » (§ 182).

Enfin, prescrivent les juges, « aux fins de garantir, en pratique, que les modèles et les critères préétablis, l'usage qui en est fait ainsi que les bases de données utilisées ne présentent pas un caractère discriminatoire et soient limités au strict nécessaire au regard de l'objectif de prévenir des activités de terrorisme présentant une menace grave pour la sécurité nationale, la fiabilité et l'actualité de ces modèles et de ces critères préétablis ainsi que des bases de données utilisées doivent faire l'objet d'un réexamen régulier » (§ 182). Le second arrêt de la Cour entend préciser encore ces limites à l'utilisation de systèmes d'apprentissage automatique.

## L'analyse des données PNR par des outils d'IA aux fins d'identifier les terroristes et les criminels

Dans cette seconde affaire, la Cour de justice était saisie d'un recours visant à faire constater l'invalidité de certaines dispositions de la directive PNR, sur base de leur contrariété alléguée à la Charte des droits fondamentaux de l'UE. Plus précisément, une association de défense des libertés remettait en question la transposition belge de la directive PNR devant la Cour constitutionnelle du même pays, ce qui a conduit ladite Cour à interroger les juges européens quant à l'interprétation et à la validité de la directive PNR elle-même, vis-à-vis des droits fondamentaux au respect de la vie privée et familiale, et à la protection des données à caractère personnel.

En effet, la directive PNR prévoit, en son article 6, que des données relatives aux passagers aériens (nom, itinéraire, dates de voyage, coordonnées, modes de paiement, informations relatives aux bagages, etc.), recueillies par les transporteurs aériens, doivent systématiquement être communiquées aux autorités publiques compétentes des États Membres. Ces données sont ensuite confrontées « aux bases de données utiles aux fins de la prévention et de la détection des infractions terroristes et des formes graves de criminalité ainsi que des enquêtes et des poursuites en la matière [...] ; ou [traitées] au regard de critères préétablis ». Dans ce cas, la directive prévoit que « [l]'évaluation des passagers [...] au regard de critères préétablis est réalisée de façon non discriminatoire. Ces critères préétablis [...] ciblés, proportionnés et spécifiques [...] ne sont en aucun cas fondés sur l'origine raciale ou ethnique d'une personne, ses opinions po-

litiques, sa religion ou ses convictions philosophiques, son appartenance à un syndicat, son état de santé, sa vie sexuelle ou son orientation sexuelle ».

Dans sa décision, la Cour de justice relève tout d'abord que « la directive PNR comporte des ingérences d'une gravité certaine dans les droits [fondamentaux à la vie privée et à la protection des données à caractère personnel], dans la mesure notamment où elle vise à instaurer un régime de surveillance continu, non ciblé et systématique, incluant l'évaluation automatisée de données à caractère personnel de l'ensemble des personnes faisant usage de services de transport aérien » (§ 111). Partant de ce constat, la Cour rappelle, dans cette affaire également, les principes de légalité et de proportionnalité et en examine le respect par la directive en cause, à savoir son aptitude à atteindre les objectifs légitimes poursuivis, et la stricte nécessité des ingérences imposées pour y parvenir.

À défaut d'analyser ici l'ensemble des considérations qui ont mené la Cour à rendre sa décision, notons toutefois qu'au cours de cet examen des mesures imposées par la directive PNR, la Cour fournit une interprétation restrictive des dispositions de la directive afin de conclure à sa validité par rapport à la Charte des droits fondamentaux de l'UE. Au-delà, elle entend limiter les usages qui peuvent être faits de l'IA dans le cadre des contrôles opérés par les autorités publiques. Ce faisant, la Cour pose une série de jalons qui conditionnent, voire limitent, l'usage d'outils d'IA par les autorités publiques dans le cadre de l'application des dispositions de la directive PNR.

Ainsi, lorsqu'elle analyse la nécessité des ingérences imposées, la Cour indique notamment que les analyses automatisées des données PNR présentent un taux d'erreur important, car elles sont basées sur des données non vérifiées et sur des modèles et critères préétablis. Les juges européens notent, à cet égard, qu'en 2018 et 2019, cinq personnes sur six identifiées par des moyens automatisés comme présentant un risque élevé ont ultérieurement été considérées comme des concordances positives erronées lors d'un réexamen par des moyens non automatisés. Partant, la Cour insiste sur le fait qu'aucune décision produisant des effets préjudiciables significatifs à l'égard d'une personne ne peut être prise sur le seul fondement d'un traitement automatisé des données PNR. Un traitement ultérieur de ces données, par des moyens non automatisés, est requis pour valider ou infirmer une concordance positive établie par un outil informatique.

Ensuite, s'agissant des « bases de données utiles » auxquelles les données personnelles des voyageurs peuvent être confrontées par les autorités publiques, la Cour apporte plusieurs précisions. Tout d'abord, elle limite sévèrement les bases de données susceptibles d'être utilisées dans le cadre de cette investigation. En premier lieu, la Cour indique qu'il s'agit des seules bases de données « concernant les personnes ou les objets recherchés ou faisant l'objet d'un signalement, conformément aux règles nationales, internationales et de l'Union applicables à de telles bases de données » (§ 187). En second lieu, la Cour détermine que ces bases de données doivent être exploitées « en rapport avec la lutte contre des infractions terroristes et des formes graves de criminalité présentant un lien objectif [. . . ] avec le transport aérien des passagers » (§ 191). Enfin, la Cour note que ces bases de données utiles doivent être gérées ou exploitées par des autorités publiques compétentes, dans le cadre de leur mission de lutte contre le terrorisme et les formes graves de criminalité. Or, ces autorités doivent être désignées de manière limitative par les États membres en application de la directive PNR.

En outre, s'agissant cette fois du traitement des données des passagers « au regard de

critères préétablis », la Cour suit les conclusions de son Avocat Général [5] et prend position contre l'utilisation, par les autorités publiques, d'outils d'IA fonctionnant sur base d'apprentissage machine, dès lors que ceux-ci ont la capacité de modifier, de manière autonome, le processus de l'évaluation des passagers. En particulier, les systèmes d'IA susceptibles de modifier les critères d'évaluation ou encore leur pondération sont prohibés, puisque de telles modifications seraient contraires au caractère préétabli desdits critères. De ce fait, seuls les outils fonctionnant grâce à des règles et pondérations entièrement préétablies par des humains – et sans capacité d'adaptation autonome ultérieure –, soit les seuls systèmes experts d'IA qualifiée de symbolique, à l'exclusion des systèmes d'apprentissage machine, pourraient être mis à contribution dans le cadre des contrôles permis par la directive PNR. La Cour ajoute, pour le surplus, que le recours aux technologies d'apprentissage machine pourrait priver d'effet utile le réexamen obligatoire, mentionné ci-avant, des concordances positives par des moyens non automatisés. En effet, « compte tenu de l'opacité caractérisant le fonctionnement des technologies d'intelligence artificielle, il peut s'avérer impossible de comprendre la raison pour laquelle un programme donné est parvenu à une concordance positive » (§ 195). Dans le même ordre d'idées, la Cour ajoute que l'utilisation de technologies d'IA fonctionnant sur base de l'apprentissage machine serait « susceptible de priver les personnes concernées également de leur droit à un recours juridictionnel effectif [. . . ], en particulier pour contester le caractère non discriminatoire des résultats obtenus » (§ 195). Notons certes que, l'existence d'une concordance positive établie par un système fonctionnant grâce à l'apprentissage machine n'empêcherait pas le réexamen ultérieur, par un agent humain, de l'ensemble du dossier d'un passager, sans tenir compte des facteurs analysés par machine (ou de leur poids), afin de prendre une décision finale de maintien – ou de suppression – de la concordance positive. En revanche, l'argument fondé sur la difficulté pour les passagers d'accéder à un recours juridictionnel effectif doit être salué : les passagers aériens seraient en effet dans l'incapacité de contester le caractère non discriminatoire des facteurs utilisés par les systèmes d'IA d'apprentissage machine : l'opacité de ces systèmes empêche, de fait, les personnes concernées de comprendre quels facteurs sont pris en compte pour établir une concordance positive automatisée, et comment ceux-ci sont mis en œuvre.

Enfin, mentionnons qu'en vertu de l'article 6 de la directive PNR, les « critères préétablis » à l'aune desquels les données des voyageurs sont évaluées doivent faire l'objet d'un réexamen régulier. A ce sujet, la Cour de justice apporte également des précisions. Elle indique ainsi que, dans le cadre de ce réexamen, les critères retenus doivent être actualisés en tenant particulièrement compte de l'expérience acquise dans le cadre de leur application, de manière à réduire autant que possible le nombre (fort élevé) de résultats de type « faux positifs ». Cette manière de procéder, dit la Cour, doit contribuer au caractère strictement nécessaire de l'application de ces critères – et donc justifier la stricte nécessité des ingérences dans les droits fondamentaux imposées en vertu de la directive PNR.

## Conclusion

Ce rapide survol des décisions rendues par les juges européens, en matière d'usage d'outils d'intelligence artificielle par les autorités publiques en charge de la sécurité publique, démontre que lesdits juges dessinent le cadre dans lequel une police « algorithmique », respectueuse des droits fondamentaux des individus et tenant compte des risques accrus engendrés

par l'IA, devra se développer.

Dans ce contexte, d'aucuns pourraient s'interroger sur la transposition de certaines parties du raisonnement de la Cour, dans ces deux affaires, à d'autres domaines, y compris dans le secteur privé, dans lesquels des outils d'IA sont utilisés pour prendre des décisions qui ont un impact significatif sur les individus, comme des outils d'octroi de crédit, de mesure de l'assurabilité ou de l'employabilité des personnes. L'interdiction d'utiliser des outils d'IA fonctionnant sur base d'apprentissage machine ne pourrait-elle pas recevoir une portée plus large, dès lors que leur utilisation risque de priver les personnes concernées également de leur droit à un recours juridictionnel effectif, du fait de l'opacité de ces outils ?

## Références

[1] *Directive 2002/58 du Parlement européen et du Conseil du 12 juillet 2002 concernant le traitement des données à caractère personnel et la protection de la vie privée dans le secteur des communications électroniques (dite directive « Vie privée et communications électroniques »), OJ L 201, 31 juillet 2002.*

[2] *Directive 2016/680 du Parlement européen et du Conseil du 27 avril 2016 relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel par les autorités compétentes à des fins de prévention et de détection des infractions pénales, d'enquêtes et de poursuites en la matière ou d'exécution de sanctions pénales, et à la libre circulation de ces données, et abrogeant la décision-cadre 2008/977/JAI du Conseil (dite directive « Police-Justice »), OJ L 119, 4 mai 2016.*

[3] *Directive 2016/681 du Parlement européen et du Conseil du 27 avril 2016 relative à l'utilisation des données des dossiers passagers (PNR) pour la prévention et la détection des infractions terroristes et des formes graves de criminalité, ainsi que pour les enquêtes et les poursuites en la matière (dite directive « PNR »), OJ L 119, 4 mai 2016.*

[4] *CJ, arrêts Privacy International, La Quadrature du Net e.a., French Data Network e.a., et Ordre des barreaux francophones et germanophone, affaires C-623/17, C-511/18, C-512/18 et C-520/18*, 2020.

[5] *Av. gén. M. G. Pitruzzella, concl. préc. CJ, arrêt Ligue des droits humains c. Conseil des ministres, affaire C-817/19*, 2022.

[6] *CJ, arrêt Ligue des droits humains c. Conseil des ministres, affaire C-817/19*, 2022.

AfIA
Association française
pour l'Intelligence Artificielle

# Le développement, le déploiement et l'utilisation d'un système d'arme létale autonome dans un conflit armé : légalité et responsabilité

Par

**Alyson BERRENDORF**
*Aspirante F.R.S-FNRS*
*Université de Liège, Belgique*
*alyson.berrendorf@uliege.be*

**Noémi BONTRIDDER**
*Chercheuse au CRIDS / NaDIS*
*Université de Namur, Belgique*
*noemi.bontridder@unamur.be*

## Introduction

Mettant à profit les avancées dans le domaine de l'intelligence artificielle, la recherche dans le secteur militaire est orientée vers une intégration croissante de l'autonomie dans les armes létales. Les « systèmes d'armes létales autonomes » (ci-après, « SALA ») ainsi en cours de création ou déjà créés (selon la définition choisie), ont pour caractéristique d'exempter partiellement ou totalement les humains des opérations de sélection et d'engagement des cibles sur le champ de bataille. En effet, après son activation ou déclenchement initial par un opérateur humain, c'est le système lui-même – usant de ses capteurs, de sa programmation informatique (*software*) et de son armement [10] – qui identifie, sélectionne et engage – ou en d'autres termes, libère la force envers – la cible. D'autres expressions sont fréquemment utilisées pour désigner ces systèmes, telles que « systèmes intelligents », « robots soldats », « robots tueurs » ou « drones autonomes », celles-ci dénotant une approche favorable ou non à leur égard.

Cette autonomisation emporte un nombre considérable d'enjeux politiques, juridiques et éthiques. C'est pourquoi depuis 2014, des discussions sur ces nouvelles armes ont lieu dans les instances onusiennes et un groupe d'experts gouvernementaux spécialisé (ci-après, « GEG ») fut institué. Parmi les enjeux discutés, deux problématiques nous occupent dans le cadre de nos recherches sur le sujet, que la présente contribution entend synthétiser. La première concerne le degré de contrôle que doit conserver l'humain sur l'usage de la force pour être à même de respecter le droit international des conflits armés. La deuxième concerne la responsabilité imputable aux différents acteurs impliqués dans le développement, le déploiement et l'utilisation d'un SALA.

## Le respect des règles relatives à la conduite des hostilités

### Les destinataires des règles relatives à la conduite des hostilités

Le droit international des conflits armés (ou *droit international humanitaire*) peut s'appliquer à tous les sujets du droit international : aux individus, aux États, aux organisations internationales, aux mouvements de libération nationale et autres collectivités paraétatiques, et aux parties aux prises dans un conflit armé non international. Bien que l'éventualité de l'octroi de la personnalité légale aux robots ait été envisagée en 2017 par le Parlement européen [2] ainsi que, subséquemment, par le GEG [3], celle-ci a bien vite été abandonnée au vu de l'incohérence conceptuelle et l'inutilité pratique d'un tel octroi. Par consé-

quent, les robots – ou plus généralement les systèmes – ne peuvent être titulaires de droits ni d'obligations, et seuls les humains qui créent, déploient et utilisent ceux-ci sont tenus de respecter le droit applicable.

En l'occurrence, en l'état, le droit des conflits armés [1] habilite les seuls combattants humains à « lancer une attaque », à l'exclusion des machines. À considérer qu'il y a « attaque » dès qu'une personne ou un objet est mis « directement en danger », le déploiement d'un système susceptible d'engager une cible constitue en tant que tel le lancement d'une attaque [7]. Puisque les règles relatives à la conduite des hostilités s'adressent à ceux qui planifient, décident et effectuent une attaque, le combattant humain est tenu de respecter les principes de distinction, de proportionnalité et de précaution, quel que soit le niveau d'autonomie de l'arme dont il se sert. Ainsi, il doit s'assurer que l'arme qu'il utilise ne l'empêche pas de procéder aux évaluations légales requises par ces principes.

### Le principe de distinction

Le principe de distinction contient deux composantes. Premièrement, les parties au conflit doivent en tout temps faire la distinction entre civils et combattants, et ne peuvent diriger leurs attaques que contre des combattants ou contre des civils qui participent directement aux hostilités [45]. Deuxièmement, les parties au conflit doivent en tout temps faire la distinction entre les biens de caractère civil et les objectifs militaires, et ne peuvent diriger leurs attaques que contre des objectifs militaires [46]. Le respect de ce principe nécessite une évaluation circonstanciée de la situation au moment de l'attaque, attaque que le combattant ne peut lancer qu'en ayant visé et en sachant ce qu'il vise, dans le respect des règles relatives à la conduite des hostilités. Le combattant ne peut donc déléguer à un système l'opération de sélection de la cible ni la décision d'engagement de celle-ci, s'il ne s'est pas préalablement assuré de toutes les mesures de prévisibilité et de fiabilité quant à la cible qui sera ainsi engagée.

### Le principe de proportionnalité

Le principe de proportionnalité « interdit de lancer des attaques dont on peut attendre qu'elles causent incidemment des pertes en vies humaines dans la population civile, des blessures aux personnes civiles, des dommages aux biens de caractère civil, ou une combinaison de ces pertes et dommages, qui seraient *excessifs* par rapport à l'avantage militaire concret et direct attendu » [47]. Cette évaluation dépend non seulement des circonstances de l'espèce, mais aussi du point de vue de la personne qui fait cette évaluation [15]. Malgré de nombreuses discussions, les États ne se sont pas accordés sur la signification des éléments devant être mis en balance pour procéder à l'évaluation de la proportionnalité de l'attaque. Quoi qu'il en soit, la proportionnalité d'une attaque n'est en tant que telle pas formulable à l'avance, et par conséquent il nous semble peu probable qu'il soit un jour possible de programmer une arme afin qu'elle procède à l'évaluation nécessaire, encore moins de manière prévisible.

### Le principe de précaution

Les belligérants sont en outre tenus de prendre constamment toutes les mesures de précaution pratiquement possibles pour éviter que les principes de distinction et de propor-

45. DIH coutumier, règles 6 et 7 ; Protocole I, articles 48, 51, §§ 2-3 et 52.
46. DIH coutumier, règle 7 ; Protocole I, articles 48 et 52.
47. Nous soulignons ; DIH coutumier, règle 14 ; Protocole I, article 51, § 5, b).
48. DIH coutumier, règles 15 à 21 ; Protocole I, article 57.

tionnalité ne soient violés et pour réduire l'impact de l'attaque sur la population civile [48]. Les commentateurs des protocoles additionnels aux Conventions de Genève [13] ont à ce propos relevé que l'interprétation de ce que « tout ce qui est pratiquement possible » implique « sera une question de bon sens et de bonne foi ». Or, le bon sens et la bonne foi sont des qualités humaines [5], et le GEG a conclu que les systèmes d'armes en cours de développement ne doivent pas être considérés comme dotés d'attributs humains [4]. Ce sont donc les humains qui doivent prendre les mesures de précaution prescrites, éventuellement en faisant usage de technologies avancées permettant de renforcer le respect du principe.

## La responsabilité imputable

### Deux régimes de responsabilité

Parmi diverses objections au développement des SALA, celle ayant trait au gap de responsabilité (*accountability gap*) est sans doute, d'un point de vue légal, la plus préoccupante : « *The premise of this objection is that an AWS [(SALA en français)] is capable of 'making its own decisions', thereby determining its own behaviour. Consequently, no person or organisation could be held accountable for the actions of the weapon system, due to a lack of control over, or perhaps even knowledge of, those decisions and actions* » [12]. Beaucoup se sont dès lors inquiétés de ce vide juridique dans le domaine répressif, en indiquant que toute tentative pour le combler serait nécessairement vouée à l'échec, notamment à cause de la nature intrinsèquement imprévisible de ces systèmes. Pour autant, bien que la question soit complexe, des pistes de réflexion demeurent à notre disposition : il convient de distinguer, d'une part, la responsabilité étatique et, d'autre part, la responsabilité des individus, toutes deux sous-tendues par des considérations différentes, mais complémentaires. La première répond à un cadre de responsabilité de nature collective : elle entend fournir un régime de responsabilité « *for any act or omission that would constitute a breach of a state's international obligations, and they cover the conduct of any agents whose acts are attributable to the state* ». La seconde, quant à elle, assure « *an individualized form of accountability for certain serious violations of IHL* » [6].

### Responsabilité étatique

Malgré plusieurs controverses doctrinales, la responsabilité étatique ne semble pas nécessairement affectée par l'autonomisation des armes létales. Si le déploiement d'un SALA engendre un acte contraire aux règles relatives à la conduite des hostilités, le déploiement de cette arme par les membres des forces armées de l'État doit être considéré comme une violation du droit pour que l'État puisse être tenu responsable des dégâts causés et donc tenu à réparation. Les propos développés dans la section précédente permettent une telle imputabilité. À défaut, la spécificité du perfectionnement des SALA doit nous rendre vigilants quant au lien de causalité entre le comportement humain attribuable à l'État et la violation de la norme. Nous notons que la responsabilité de l'État peut ainsi être engagée pour l'utilisation d'un SALA par ses forces armées sous certaines conditions, mais qu'elle peut aussi l'être aux stades antérieurs du développement et de l'approvisionnement.

### Responsabilité individuelle

Plus ardue est la problématique relative à la responsabilité pénale individuelle, et ce, principalement en raison de trois éléments : (1) le fonctionnement spécifique, voire inédit, de ce type d'arme intégrant des caractéristiques autonomes dans des tâches critiques (de sélec-

tion et d'attaque) ; (2) la conceptualisation de la responsabilité pénale attachée à l'utilisation de ces armes au travers de l'élément moral de l'intentionnalité ; et (3) le problème du « *many hand* ».

Concernant le premier versant, il fut avancé que l'attribution de la responsabilité pénale, selon la structure de notre système répressif, était inadéquate afin de répondre aux menaces que représentent les SALA [11]. En effet, de par la définition qui est retenue de ces systèmes, ces derniers seraient capables de sélectionner et engager des cibles, sans intervention humaine. Se pose alors la question de la qualification juridique qui doit être associée à ces armes, sur base de leurs caractéristiques techniques, en tant qu'auteur de l'infraction, instrument de l'infraction, ou encore comme appartenant à une nouvelle catégorie en devenir – ce qui aura une influence sur le système répressif choisi.

Le deuxième versant a trait à l'intentionnalité que requiert notre droit répressif : le niveau d'imprévisibilité de ces systèmes apporte une dimension inédite en comparaison à toutes les problématiques étudiées jusqu'alors. En effet, si le système fonctionne comme prévu, mais qu'il a été développé ou utilisé par un individu dont le dessein est d'attaquer des personnes ou des biens protégés par le droit international humanitaire, la responsabilité humaine corrélative est facilement attribuable, étant donné que l'élément moral du crime de guerre est présent. Toute autre est la question d'un système fonctionnant comme prévu sur le champ de bataille, mais dont les conditions entourant son déploiement sont changeantes, dynamiques, complexes, et par nature, différentes de celles exercées dans un laboratoire. Dans ce cas, le déploiement d'un SALA pourrait entraîner une violation du droit international humanitaire « *without anyone acting intentionally* » [9, 8]. Cette manière de procéder nous oblige à repenser les normes de la responsabilité pénale attachées à ces comportements. Les SALA relancent les débats relatifs à la répression des comportements à risque, en particulier la question de savoir si le fait d'exécuter un comportement risqué quant aux effets d'une attaque pourrait ou non entraîner une responsabilité pénale internationale pour crime de guerre [6].

Le dernier versant, qui met en lumière le « *many hands problem* » est pour le moins inédit. Le processus décisionnel relatif au déploiement de la force létale serait à rattacher à l'opérateur humain en charge de l'utilisation du SALA, ce qui, corrélativement, entraînerait sa responsabilité pénale en cas d'infraction. Pour autant, cette affirmation semble perdre de vue que l'opérateur humain n'est pas le seul à être impliqué dans le processus de ciblage et d'attaque. Sur le champ de bataille, au côté de l'opérateur humain, le commandant (ou le supérieur hiérarchique), ainsi que le *Legal Advisor* jouent un rôle fondamental dans la détermination de la légalité d'une attaque lors de la conduite des hostilités [14]. En amont de ces intervenants, les rôles du programmeur des lignes de code intégrées dans le SALA afin de se soumettre aux principes du DIH décrits ci-dessus, et du producteur de ce même système ne doivent pas être oubliés de la chaîne de responsabilité.

Ainsi, il sera question d'analyser l'implication dans le processus décisionnel de chacun des humains pouvant y avoir joué un rôle et de voir, en cas d'infraction, comment la responsabilité individuelle peut leur être imputée. Bien que ce processus soit complexe, cela ne signifie pas pour autant qu'aucun être humain ne puisse être juridiquement tenu responsable des actions d'un SALA. C'est pourquoi, à notre estime, bien que l'attribution d'une responsabilité pénale en matière de SALA nous impose de repenser notre système répressif, particulièrement au regard de l'élément moral, le postulat

d'un « *accountability gap* » pour le développement, le déploiement et l'utilisation d'un SALA est inexact.

## Références

[1] *Protocole additionnel aux Conventions de Genève du 12 août 1949 relatif à la protection des victimes des conflits armés internationaux (Protocole I), signé à Genève le 8 juin 1977, entré en vigueur le 7 décembre 1978*. Article 43, §§ 1-2., 1978.

[2] *Résolution du Parlement européen du 16 février 2017 contenant des recommandations à la Commission concernant des règles de droit civil sur la robotique*. 2015/2103(INL), point 59, f), 2017.

[3] *Rapport du Groupe d'experts gouvernementaux sur les systèmes d'armes létaux autonomes sur sa session de 2017*. Annexe II – Résumé des débats, ONU CCW/GGE.1/2017/3, p. 13, par. 53, 2017.

[4] *Rapport du Groupe d'experts gouvernementaux sur les technologies émergentes dans le domaine des systèmes d'armes létaux autonomes sur sa session de 2019*. ONU CCW/GGE.1/2019/3, p. 15, principe directeur i), 2019.

[5] D. Akerson. The illegality of offensive lethal autonomy. In D. Saxon, editor, *International humanitarian law and the changing technology of war*, page 81. Martinus Nijhoff Publishers, 2013.

[6] M. Bo, L. Bruun, and V. Boulanin. *Retaining Human Responsibility in the Development and Use of Autonomous Weapon Systems : On Accountability for Violations of International Humanitarian Law Involving AWS*. SIPRI, 2022.

[7] V. Boulanin, N. Davison, N. Goussac, and M. Peldán Carlsson. Limits on autonomy in weapon systems. Technical report, SIPRI, p. 6, 2020.

[8] V. Chiappini Koscina. Prosecuting killer robots : Allocating criminal responsabilities for grave breaches of international humanitarian law committed by lethal autonomous weapon systems. In B. Custers et E. Fosch-Villaronga, editor, *Law and Artificial Intelligence, Information Technology and Law Series*, volume 35, page 154. 2022.

[9] R. Crootof. War torts : Accountability for autonomous weapons. *NYU Law Review*, page 1366, 2016.

[10] N. Davison. A legal perspective : Autonomous weapon systems under international humanitarian law. *UNODA Occasional Papers – Perspectives on Lethal Autonomous Weapon Systems*, 30 :6, 2017.

[11] Human Rights Watch & IHRC. Mind the gap : The lack of accountability for killer robots. Technical report, p. 6, 2015.

[12] T. McFarland. *Autonomous Weapon Systems and the Law of Armed Conflict*. Cambridge University Press, p. 8, 2020.

[13] C. Pilloud and J. De Preux. Article 57 – précautions dans l'attaque. In C. Swinarski et B. Zimmerman Y. Sandoz, editor, *Commentaire des protocoles additionnels du 8 juin 1977 aux Conventions de Genève du 12 août 1949*, pages 700, par. 2198. Martinus Nijhoff Publishers, 1986.

[14] N. Sharkey. Automating warfare : Lessons learned from the drones. *Journal of Law, Information and Science*, 21(2) :7, 2012.

[15] T.P.I.Y. *Final Report to the Prosecutor by the Committee Established to Review the NATO Bombing Campaign Against the Federal Republic of Yugoslavia*. 13 juin 2000, par. 50.

AfIA

Association française
pour l'Intelligence Artificielle

# Comptes rendus de journées, événements et conférences

AfIA
Association française pour l'Intelligence Artificielle

# Le défi de l'AfIA pour la Nuit de l'Info 2022

Par

**Florence BANNAY**
*IRIT / ADRIA*
*Université de Toulouse*
*florence.bannay@irit.fr*

**Anne-Gwen BOSSER**
*Lab-STICC / COMMEDIA*
*École Nationale d'Ingénieurs de Brest*
*bosser@enib.fr*

**Maëlic NEAU**
*maelic.neau@enib.fr*

La Nuit de l'Info 2022 a eu lieu de 16h35 à 8h00 dans la nuit du 1 au 2 décembre sur le thème SexInfo La prévention par le jeu, il s'agissait de développer une application servant à diffuser des informations sur la sexualité au moyen d'un *serious game* en ligne (jeu vidéo, escape game, jeu de rôle textuel, quiz, jeu de cartes contre une IA, etc.) L'AfIA a proposé pour la 7e année un défi « Mettez de l'intelligence dans votre moteur ». Pour l'édition 2022, pour mieux prendre en compte l'importance grandissante de l'IA dans la vie de tous les jours et dans la formation des jeunes informaticiennes et informaticiens, et au vu des soumissions 2021, le Collège CECILIA a proposé d'ajouter les considérations éthiques dans l'évaluation de la solution proposée.

Le principe de la Nuit de l'Info est simple : *Le jeudi 1, au coucher du Soleil, en séance plénière : les organisateurs remettent un sujet (le même pour toute la France) aux participants. Les étudiants s'organisent en groupes (sur un ou plusieurs sites) : ils développent un projet (informatique, marketing, rédactionnel, etc.) tout en ciblant un ou plusieurs défis. Le vendredi 2, au lever du Soleil, les développements sont figés. Le vendredi matin, pendant que les étudiants dorment, des jurys se réunissent (un jury par défi) et examinent les travaux réalisés par les différentes équipes durant la nuit.*

## Le défi de l'AfIA

Le défi qu'a proposé l'AfIA s'intitulait « Mettez de l'intelligence dans votre moteur », il a été proposé et organisé par le collège Création d'Événements Collaboratifs, Inclusifs et Ludiques en IA (CECILIA), qui défend l'apprentissage de l'IA grâce à la pratique coopérative et l'expérimentation. Le jury était constitué de : Florence BANNAY, Anne-Gwenn BOSSER, et Maëlic NEAU. Voici sa description : *Vous mettez en œuvre une ou plusieurs méthodes d'Intelligence Artificielle (IA) dans votre projet et vous indiquez en quoi ces méthodes rendent votre réalisation plus performante ou pertinente. Vous pouvez faire appel à des techniques classiques, en cours de développement ou futuristes. L'équipe qui aura mis le plus en avant les avantages de l'utilisation de l'IA dans son projet remportera ce défi.*

## Éléments attendus/critères de notation

- Description des problématiques IA,
- Explication de l'intégration de l'IA :
  - décrivez ce que vous avez réalisé qui relève de l'IA ou de ce qui aurait pu être fait avec des outils d'IA existants ou imaginaires,
  - différenciez les bouts de codes et logiciels existants (citez vos sources), des parties spécifiées et codées en propre pour le défi,
  - un « use-case » / exemple illustratif,
  - fournissez une annexe technique
- Analyse des avantages et inconvénients

Cinq notes ont été données : Applicabilité / mise en œuvre, Innovation, IA (évaluation qualitative et quantitative : richesse de la solution proposée, spectre couvert, éthique de la solution proposée), Qualité des explications.

### Résultats

Parmi les 19 équipes inscrites, le jury souhaite féliciter les 11 équipes qui ont réussi à aboutir à la remise d'un projet dans le temps imparti. Il s'agit des équipes :

- « !Win »
- « Invincible_ISITcom »
- « Terminator_Genisys_ISITcom »
- « Bruh »
- « Les Dodos Insomniaques, le retour »
- « Team Symphonie »
- « Grincheux »
- « EPI_8 shades of brains »
- « Unchained »
- « les simploniens »
- « Codelab »

### Première place ex-aequo

La première place a été attribuée aux équipes « Codelab » et « Les Dodos Insomniaques, le retour » qui se démarquent des autres soumissions. La solution présentée par « Codelab » est un site Web en ligne fonctionnel consistant en un quizz auquel l'utilisateur doit participer afin de ester ses connaissances sur le thème des maladies sexuellement transmissibles. Le site affiche des statistiques sur le genre et l'âge des personnes les mieux informées sur le VIH. L'utilisation de GPT-3 et Dall-E est pertinente et bien intégrée. L'ajout des requêtes faites vers GPT-3 sur le GitHub est un plus pour la ré-utilisabilité et la transparence du code. Notons cependant que la diversité entre les propositions positives et négatives au quiz n'est pas suffisante, c'est un biais assez commun de GPT-3 qui peut être amélioré en changeant les paramètres de génération. Attention à l'usage de Dall-E, certaines images ne sont pas vraiment en accord avec le propos.

L'équipe des « Dodos insomniaques, le retour » a proposé une approche ludique, même s'il ne s'agit pas d'un *serious game* à proprement parler. Trois bibliothèques d'IA sont utilisées : Dall-E pour que les utilisateurs génèrent des images de Mèmes sur l'information au SIDA, un classificateur pour modérer automatiquement des phrases jugées offensantes. Notons qu'il aurait peut-être été souhaitable de développer des précautions à ce sujet : puisqu'il s'agirait de Mèmes qui seraient estampillés du sceau de l'association. La curation des contenus générés est à ce jour un problème difficile à régler sans intervention humaine. Le troisième outil est un LSTM entraîné sur un jeu de données public (*Kaggle*) pour générer une prédiction de l'espérance de vie des malades.

Le jury a considéré que les deux équipes méritaient un prix, chacune dans sa catégorie : « Les Dodos insomniaques, le retour » proposent une solution un peu moins bien adaptée au sujet national, mais touchant un large spectre de l'IA et avec des explications de qualité. L'équipe « Codelab » a fourni une solution qui répond bien au défi national avec une bonne applicabilité, mais pour laquelle les explications ne sont pas aussi bien détaillées.

### Les 9 autres équipes

Nous avons eu beaucoup de propositions ayant peu de rapport avec le sujet national ou peu de rapport avec l'IA. Ainsi, l'équipe « !Win » qui a mis en œuvre un jeu narratif dont vous êtes le héros permettant de mettre l'utilisateur en situation et de l'amener en fin de scénario à des liens vers de l'information sur le sujet, malgré cette bonne idée de base, aucune IA n'est utilisée ou suggérée pour rendre le jeu intéressant/efficace/« intelligent ».

### Bilan

Le jury a décidé d'attribuer la récompense de 200 euros à chacune des 2 équipes « Codelab » et « Les Dodos Insomniaques, le retour ». Merci à toutes les équipes participantes, félicitations à l'équipe victorieuse, et rendez-vous l'année prochaine !

AfIA
Association française
pour l'Intelligence Artificielle

# Thèses et HDR du trimestre

Si vous êtes au courant de la programmation de soutenances de thèses ou HDR en Intelligence Artificielle cette année, vous pouvez nous les signaler en écrivant à redaction@afia.asso.fr.

## Thèses de Doctorat

**Razanne Abu AISHEH**
« Context-Aware Information Gathering and Processing Towards Supporting Autonomous Systems in Industry 4.0 Scenarios »
Supervision : *Thomas WATTEYNE*
*Francesco BRONZINO*
Le 27/02/2023, à Sorbonne universite

**Alaa ZREIK**
« Semantic trajectory analysis for the prediction of the physical state of the collections at the BnF »
Supervision : *Zoubida KEDAD*
Le 16/01/2023, à l'Université Paris-Saclay

**Jiayi HONG**
« Machine Learning Supported Interactive Visualization of Hybrid 3D and 2D Data for the Example of Plant Cell Lineage Specification »
Supervision : *Tobias ISENBERG*
*Alain TRUBUIL*
Le 14/02/2023, à l'Université Paris-Saclay

**Fabrice POPINEAU**
« Approche Logique de la Personnalisation dans les Environnements Informatiques pour l'Apprentissage Humain »
Supervision : *Nicole BIDOIT TOLLU*
Le 12/01/2023, à l'Université Paris-Saclay

**Roxane Elias MALLOUHY**
« Predictive analysis of time series in various application contexts »
Supervision : *Christophe GUYEUX*
*Chady Abou JAOUDE*
Le 05/01/2023, à l'Université de Bourgogne Franche-Comte

## Habilitations à Diriger les Recherches

Nous n'avons malheureusement pas eu connaissance ce trimestre d'HDR dans le domaine de l'IA.

N'hésitez pas à nous envoyer les informations concernant celles dont vous avez entendu parler. (redaction@afia.asso.fr).

AfIA
Association française
pour l'Intelligence Artificielle

# À PROPOS DE
## L'AfIA

L'objet de l'AfIA, Association Loi 1901 sans but lucratif, est de promouvoir et de favoriser le développement de l'Intelligence Artificielle (IA) sous ses différentes formes, de regrouper et de faire croître la communauté française en IA et, à la hauteur des forces de ses membres, d'en assurer la visibilité.

L'AfIA anime la communauté par l'organisation de grands rendez-vous. Se tient ainsi chaque été une semaine de l'IA, la Plate-forme IA (PfIA 2022 à Saint-Étienne, PfIA 2023 à Strasbourg) au sein de laquelle se tiennent la Conférence Nationale d'Intelligence Artificielle (CNIA), les Rencontres des Jeunes Chercheurs en IA (RJCIA) et la Conférence sur les Applications Pratiques de l'IA (APIA) ainsi que des conférences/journées thématiques hébergées qui évoluent d'une année à l'autre, sans récurrence obligée.

Ainsi, PfIA 2023 héberge du 3 au 7 juillet 2023 à Strasbourg, outre la 26e CNIA, les 21es RJCIA et la 9e APIA : les 6 conférences CAp, IC, JFPC, JFSMA, JIAF-JFPDA et SFC, 4 journées thématiques (ACAI, Jeux & IA, Résilience & IA, Santé & IA), et plusieurs tutoriels hébergés.

Forte du soutien de ses 374 adhérents à jour de leur cotisation en 2022, l'AfIA assure :

- le maintien d'un site Web dédié à l'IA reproduisant également les Brèves de l'IA ;
- une *journée industrielle* « Forum Industriel en IA » (FIIA 2022) ;
- une *journée recherche* « Perspectives et Défis en IA » (PDIA 2022) ;
- une *journée enseignement* « Enseignement et Formation en IA » (EFIA 2023) ;
- une « École Saisonnière en IA » (ESIA2023) ;
- la remise annuelle d'un prix de thèse en IA ;
- le soutien à 8 collèges ayant leur propre activité :
  - collège Industriel (janvier 2016) ;
  - collège Apprentissage Artificiel (janvier 2020) ;
  - collège Interaction avec l'Humain (juillet 2020) ;
  - collège Représentation et Raisonnement (avril 2017) ;
  - collège Science de l'Ingénierie des Connaissances (avril 2016) ;
  - collège Systèmes Multi-Agents et Agents Autonomes (janvier 2017) ;
  - collège Technologies du Langage Humain (juillet 2019) ;
  - collège Création d'Événements Collaboratifs, Inclusifs et Ludiques en IA (octobre 2021) ;
- la parution trimestrielle des Bulletins de l'AfIA ;
- un lien entre ses membres et sympathisants sur les réseaux sociaux LinkedIn, Facebook et Twitter ;
- le parrainage scientifique, mais aussi éventuellement financier, d'événements en IA ;
- la diffusion mensuelle de Brèves sur les actualités de l'IA en France (abonnement ou envoi à la liste) ;
- la réponse aux consultations officielles ou officieuses (Ministères, Missions, Organismes) ;
- la réponse aux questions de la presse, écrite ou orale, également sur internet ;
- la divulgation d'offres de collaborations, de formations, d'emploi, de thèses et de stages.

L'AfIA organise aussi des journées communes avec d'autres associations. Pour 2022 : EIAH & IA avec l'ATIEF; IoT & IA avec l'IMT; Résilience & IA avec la région ARA; Réalité Virtuelle & IA avec le GDR IG-RV; Santé & IA avec l'AIM; Simulation & IA avec le réseau DEVS/RED.

Enfin, l'AfIA encourage la participation de ses membres aux grands événements de l'IA, dont PfIA. Ainsi, les membres de l'AfIA, pour leur inscription à PfIA, bénéficient d'une réduction équivalente à deux fois le coût de leur adhésion, leur permettant d'assister à PfIA 2023 sur 5 jours au tarif de 114 € TTC !

Rejoignez-nous vous aussi et adhérez à l'AfIA pour contribuer au développement de l'IA en France. L'adhésion peut être individuelle ou au titre de personne morale. Merci également de susciter de telles adhésions en diffusant ce document autour de vous !

AfIA
Association française
pour l'Intelligence Artificielle



## Pour contacter l'AfIA


**Président**
Benoit LE BLANC
École Nationale Supérieure de Cognitique
Bordeaux-INP
109 avenue Roul, 33400 Talence
Tél. : +33 (0) 5 57 00 67 00
president@afia.asso.fr


**Serveur WEB**
http://www.afia.asso.fr

**Adhésions, liens avec les adhérents**
Isabelle SESÉ
tresorier@afia.asso.fr

## Calendrier de parution du Bulletin de l'AfIA

| | Hiver | Printemps | Été | Automne |
|---|---|---|---|---|
| Réception des contributions | 15/12 | 15/03 | 15/06 | 15/09 |
| Sortie | 31/01 | 30/04 | 31/07 | 31/10 |