IA, neurosciences et technologies

Laurent, Nathanael; Poullet, Yves; Tilman, Valerie; Fourneret, Eric; Doat, David; Guillermin, Mathieu

Published in:
Bulletin de l'AFIA

Publication date:
2023

Document Version
le PDF de l'éditeur

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.