# L'IA, l'éthique et la théorie des *baïonnettes intelligentes*

Dominique Boullier et Aurélie Jean
Pour AOC
Novembre 2024

## *Des scandales à répétition et des solutions qui tardent à venir*

Ces dernières années, les nombreux scandales et controverses algorithmiques ont mis en évidence auprès du grand public les risques inhérents à des technologies d'intelligence artificielle (IA) mal pensées, mal conçues et/ou mal utilisées [1] [2]. On peut citer le cas des premiers algorithmes de reconnaissance faciale dans les années 2010 qui ne reconnaissaient pas les visages à peau foncée [3] [4], l'algorithme test de recrutement d'*Amazon* en 2018 qui écartait systématiquement les CVs de femmes pour un poste d'ingénieur en informatique, ou encore l'algorithme de l'application *Apple Card* (développé par la banque d'affaires *Goldman Sachs*) en 2019 qui proposait des lignes de crédit jusqu'à vingt fois plus élevées pour des hommes que pour des femmes aux mêmes revenus et aux mêmes conditions fiscales. À ces cas de discrimination technologique, on peut ajouter les nombreux cas de fuites ou de mauvais usage de data comme l'affaire *Cambridge Analytica* en 2018 et l'emploi des données à caractère personnel de millions d'utilisateurs de Facebook a des fins de manipulation politique.

La plupart de ces scandales auraient pu être évités par la mise en place d'une meilleure gouvernance et le partage des leçons à tirer sur l'origine de ces erreurs [5]. Au contraire, les acteurs les plus puissants et dont les erreurs sont les plus préjudiciables, portent un discours qui tend à les déresponsabiliser. En 2024, le patron d'*OpenAI*, Sam Altman, affirme craindre ce qu'il nomme des "*désalignements sociétaux*" à l'origine selon lui des nombreux problèmes rencontrés, comme pour se désengager d'un quelconque rôle dans les tests à réaliser sur les jeux de données et les algorithmes une fois entraînés mais aussi pendant leur entraînement. De son côté, Mark Zuckerberg réaffirme régulièrement la difficulté de détecter algorithmiquement la propagation virale de fausses nouvelles alors même que la solution se trouve en grande partie dans la modification de ses algorithmes de recommandation tout en changeant de modèle économique. Ce constat est similaire, même à une moindre échelle, chez les scientifiques et ingénieurs qui écrivent et développent des algorithmes dans de nombreux domaines, et qui pour certains affirment être dépassés par ces algorithmes qu'ils ne comprennent plus. Alors même que des méthodes d'explicabilité [6] en amont et en aval du développement permettent de maîtriser en partie la logique algorithmique.

Le *AI Act* européen voté le 13 Mars 2024, constitue une brique fondamentale vers la responsabilisation des propriétaires de ces IAs, et doit être accompagné de méthodes pratiques comme la mise en place d'un nouveau paradigme en réponse au désengagement individuel et collectif : la *théorie des baïonnettes intelligentes* appliquée à l'IA.

### *Vers des intelligences artificielles dites de confiance*

L'IA est à la fois une discipline et des méthodes de calcul computationnel pour modéliser et simuler un phénomène dans l'objectif de répondre à une question ou de résoudre un problème. En pratique, un algorithme est calibré et/ou entraîné [7] sur un jeu de données représentant statistiquement les phénomènes considérés, en en capturant les signaux faibles et forts. La qualité de l'algorithme, incluant son exactitude et sa justesse, dépend de la façon de poser le problème, des pratiques de sa conception, de son développement informatique, de sa validation, de ses tests et de ses modes d'usage. On peut citer les IAs explicites et implicites (apprentissage automatique), parmi lesquelles on distingue par exemple les IAs de vision computationnelle, les IAs prédictives, ou encore les IAs génératives. Il faut également distinguer la technologie d'IA (comme un algorithme de reconnaissance d'identité sur vidéo comme celui embarqué dans les systèmes de déverrouillage de téléphone ou la douane) de son application (comme la vidéo surveillance généralisée dans les rues), afin d'en souligner les spécificités, les bénéfices et les risques.

Les propriétaires et les concepteurs doivent se responsabiliser face aux possibles erreurs intentionnelles ou accidentelles en adoptant de bonnes pratiques à chaque étape de conception d'un algorithme (de la formulation de la problématique aux informations transmises à l'utilisateur final). Par une telle gouvernance algorithmique d'excellence, on peut déployer ce qu'on nomme des IAs de confiance, dont l'exécution respecte les droits fondamentaux des citoyens par une responsabilisation de toutes les parties prenantes tout au long du processus de fabrication de l'algorithme.

### *La théorie des baïonnettes intelligentes et ses applications aujourd'hui*

La théorie des baïonnettes intelligentes a été initiée en France au cours de la seconde guerre mondiale et de l'arrêt du 10 novembre 1944, de l'affaire *Langneur* [8]. Le droit reconnaissait ainsi que toute action porteuse de dommages de la part d'une baïonnette (qui transperce un corps notamment) trouvait sa source dans une décision d'un acteur humain qui ne pouvait s'en détacher en attribuant la responsabilité à l'outil baïonnette supposé intelligent. Le soldat ne pouvait ainsi se considérer comme une simple baïonnette car il était supposé intelligent. En droit pénal, elle est un outil juridique qui

permet de condamner une personne physique qui a obéi à un ordre de sa hiérarchie, manifestement illégal. Appliquée initialement dans le domaine militaire, elle s'étend ensuite au domaine civil[1] dans la fonction publique puis dans le privé. Aux États-Unis, son équivalent américain, le *Uniform Code of Military Justice* a été mis en vigueur dans les années cinquante. Le 23 janvier 1997, l'ancien préfet Maurice Papon fut condamné au motif de "*l'illégalité d'un ordre de l'autorité légitime en matière de crime contre l'humanité [est] toujours manifeste*". Plus récemment, en 2019, des gendarmes de la ville de Nice en pleine crise des Gilets Jaunes lors d'un rassemblement, refusent d'obéir à un ordre de leur hiérarchie. Ils le jugent *disproportionné* au regard de la menace surestimée selon eux, et ont été soutenus par l'appel à cette théorie.

À l'ère de l'IA, la théorie des baïonnettes intelligentes doit être mise en relief au regard de deux dimensions ici considérées. La première dimension concerne son application dans le contexte du développement d'IAs dites autonomes, qu'elles soient à destination militaire (comme des systèmes de navigation ou des armes) ou civile (comme les véhicules autonomes). Certains domaines comme la médecine qui proscrivent l'emploi de modèles d'IA sans intervention humaine, n'ont pas à considérer cette dimension liée à l'autonomie fonctionnelle de la technologie d'IA. Néanmoins, la théorie des baïonnettes intelligentes s'applique toujours aux corps médical dans l'obéissance à leur hiérarchie. Appliqué au domaine militaire, l'autonomie des systèmes d'attaque remet en question la théorie même au regard de l'implication asymétrique des parties prenantes, avec des offenseurs et défenseurs situés parfois à des milliers de kilomètres du lieu de l'attaque, faisant perdre les unités de temps et de lieu du combat [9]. Par ailleurs, comme indiqué par Ruben Stewart et Georgia Hinds, la confiance apparente des humains vis-à-vis de la suggestion algorithmique peut entraîner un biais dit d'automatisation qui fausse leur jugement tout en les dispensant d'examiner des solutions alternatives parfois plus efficaces. La vitesse d'exécution tend également à diminuer le temps décisionnel humain alors même que la patience stratégique apporterait une perspective plus élargie de la situation.

La seconde dimension de la théorie des baïonnettes intelligentes porte sur son application directe chez les concepteurs des technologies d'IA, qui doivent se responsabiliser face aux demandes de leurs supérieurs dans un contexte de risques largement discutés et de puissance économique et politique des acteurs en pleine croissance [10]. Cela implique chez chacun de ces concepteurs de se renseigner sur les applications de la technologie qu'on leur demande de développer, ainsi que sur la manière dont la technologie est conçue en dehors de leur propre contribution. Ils doivent

[1] Selon l'article 122-4 du code pénal: "*N'est pas pénalement responsable la personne qui accomplit un acte commandé par l'autorité légitime, sauf si cet acte est manifestement illégal*"

enfin s'interroger sur les usages de la technologie en dehors de celui pour lequel elle a été initialement pensée et développée.

***La distribution des responsabilités en intelligence artificielle***

Nombreux sont ceux qui simplifient les débats sur l'IA en les réduisant quelquefois à des arguments parfois faussés et souvent à une réflexion déconnectée du terrain, ce qui empêche d'*encastrer* le sujet de l'IA dans un *design organisationnel* particulier en mobilisant un grand nombre de parties prenantes. L'anthropologue K. Polanyi parlait du *désencastrement* de l'économie [11] réalisé durant le XIXe siècle qui laissait croire que les activités d'échange économiques pouvaient être modélisées et pensées en les coupant de toute leur insertion dans des organisations, des lois, des valeurs, des conflits sociaux, etc. Translaté à l'IA, on observe que les algorithmes semblent se désencastrer de toute insertion dans une vie sociale, politique et économique complexe sans prendre en compte leur inscription dans des organisations, des lois, ou encore des pratiques. C'est par une meilleure connaissance des phases de conception et d'usage qu'on peut précisément identifier toutes les parties prenantes [12]. Même si la comparaison peut paraître simpliste dans la mesure où un système d'IA et son interaction avec son environnement (incluant l'humain) est bien plus sophistiqué qu'un rapport entre un soldat et sa baïonnette, le principe reste le même : la *distribution des responsabilités* pour lutter contre leur dilution apparente. Cette distribution permet entre autres à chaque partie prenante de définir ses responsabilités à travers la mise en évidence de ses droits, ses devoirs et ses obligations, ce qui inclut les concepteurs. Une transparence sur les procédés, les tâches constitutives de construction de l'IA en question et des informations transmises à l'utilisateur est alors un pré-requis.

Dans ce processus, les parties prenantes doivent être identifiées à travers leurs rôles et leurs responsabilités (droits, obligations, devoirs) mais aussi leurs intentions et leurs ressorts. Cette étape préalable de *distribution d'agentivité* [13] vise à définir qui fait quoi au sein de la chaîne de conception et selon quelles proportions. Cela permet à chaque partie prenante d'alerter sur un possible risque de l'IA en question en s'interrogeant sur ce qui est acceptable socialement (à l'instar de Timnit Gibru en 2020). Ou encore, en pensant aux pires scénarios, il devient fondamental de tester les algorithmes et de documenter chaque étape ainsi que le fonctionnement général de l'outil dans l'objectif de se faire comprendre de l'utilisateur final. Ainsi, on sortira de l'idée que seul l'utilisateur final est responsable, pouvant utiliser l'IA en question pour le pire comme pour le meilleur, pour au contraire affirmer que l'IA n'est pas neutre par son design [14]. Inclure les talents du métier d'application de l'IA en question permet de soutenir au mieux une telle approche en conservant un pied avec la réalité du terrain, de ses enjeux et de ses défis.

Cette distribution des responsabilités - qui s'oppose encore une fois à une quelconque dilution de ces mêmes responsabilités - permet de mitiger les risques de ces modèles comme les biais algorithmiques pouvant mener à de la discrimination technologique. Cette gouvernance algorithmique permet de réduire les risques de désencastrement des choix et des décisions réalisés tout du long du processus. L'application d'une telle gouvernance aurait certainement remis en question la décision d'Open AI de déployer massivement et publiquement ChatGPT en considérant que l'usage public fera office de banc de test, au regard des risques bien trop élevés de biais, d'erreurs issues d'hallucinations algorithmiques, de mauvaises utilisations de la part d'individus non préparés et confus, ou encore de plagiat au sein des contenus générés et autres violations de la propriété intellectuelle des données sur lesquels l'algorithme a été entraîné. La *distribution des responsabilités* au sein d'un cadre de gouvernance algorithmique permet de penser une technologie au sens moral également, au-delà de la simple dimension économique. La sociologie de l'innovation propose des méthodes pour décrire finement ces réseaux de contributeurs dans une innovation "*en train de se faire*" [15].

Le concept de *distribution des responsabilités* s'intègre correctement à la théorie de la *baïonnette intelligente* car il fait partie du vocabulaire et des concepts classiques du monde de l'internet, et parce qu'il a fait aussi l'objet d'une autre théorie en matière de sciences cognitives, la *cognition distribuée* [16]. Dans cette théorie, Edwin Hutchins montre que les processus de décision qui paraissent être des calculs très sophistiqués (comme faire le point par triangulation sur un porte-avion nucléaire) sont en réalité des effets de "*propagation d'états représentationnels grâce à des médias représentationnels*". En pratique, ce sont seulement des déplacements matériels, d'un support de communication à un autre, qui produisent un nouvel effet de connaissance. C'est ce qui se produit lorsqu'on conçoit un algorithme d'apprentissage automatique par optimisation (par rétropropagation du gradient de l'erreur), dans l'objectif d'améliorer le modèle sans réfléchir aux fondamentaux mais en ne cessant de l'améliorer en continu jusqu'à possiblement augmenter de manière significative les risques précédemment mentionnés et donner l'impression qu'on perd la maîtrise du modèle. Néanmoins, les méthodes d'explicabilité algorithmique (en constante évolution) doivent être systématiquement appliquées pour éviter l'effet boite noire et ainsi mitiger les risques.

En opposition aux industries classiques et historiques qui doivent investir dans la production d'entités tangibles nécessitant des ressources en termes de logistiques et de matières premières, l'industrie du numérique repose sur une facilité apparente d'organisation qui suppose la création ou la destruction de code informatique ou de jeux de données sans conséquences financières ou organisationnelles. En pratique, cette vision idéalisée est un frein au développement responsable qui implique des

apprentissages collectifs pour la conception d'outils fiables. En cela, le développement Agile qui suppose un processus itératif en est l'incarnation quand il est mal utilisé, c'est-à-dire de manière irresponsable.

### *Une consolidation des responsabilités individuelles pour le collectif*

Il peut en revanche sembler trop compliqué et donc coûteux de systématiquement effectuer une forme de contrôle de responsabilité permanent à chaque étape du développement ainsi que pour chaque partie prenante incluant les tierces parties. C'est pourquoi des standards reconnus de procédures de contrôle seront indispensables. Dès lors qu'une entreprise vise à se conformer à ces standards et aux labels qui lui sont associés, elle peut procéder à une *consolidation des responsabilités*. Précisons qu'aucune personne physique (incluant les développeurs) ne peut se prévaloir de cette consolidation induite par le suivi des procédures, pour se dégager de sa responsabilité individuelle en cas d'erreurs et d'incidents divers. C'est ainsi que les lanceurs d'alerte procèdent pour mettre en lumière une défaillance potentielle, volontaire ou pas.

Par exemple, lorsque Microsoft annonce en Septembre 2023 qu'il prendra à sa charge tout le contentieux lié aux infractions éventuelles de Copilot en matière de Copyright, il tend à construire une voûte de protection, qui est potentiellement coûteuse à terme mais qui crée des conditions de confiance pour les utilisateurs et pour tout le marché. Cela suppose ensuite de construire en interne la chaîne de contrôles de qualité qui deviennent dès lors indispensables, entraînant une cascade de responsabilités qui peut être inscrite dans les contrats de travail. Une telle consolidation peut stabiliser les marchés et sécuriser les enjeux juridiques, mais elle ne saurait permettre de se dérober aux contraintes d'explicabilité nécessaires pour que la responsabilité s'exerce en connaissance de cause. Elle ne constitue pas un blanc-seing pour un nouvel effet boite noire pourtant évitable - même en partie - par les nombreuses méthodes statistiques de calcul d'explicabilité.

### *La loi pour protéger les acteurs et encourager l'innovation responsable*

La théorie des baïonnettes intelligentes, levier évident pour se protéger par la responsabilité individuelle et collective, s'inscrit *de facto* dans un cadre législatif pour permettre son application. En cela, et appliquée à l'IA, cette théorie doit être soutenue - ou tout simplement rendue possible - par des textes légaux bien construits. En Europe mais aussi à travers le monde, les textes pour encadrer la collecte, le stockage et l'usage des données à caractère personnel ont pris place pour protéger les droits fondamentaux des citoyens tout en encourageant l'innovation. En Europe, le RGPD est la manifestation de cette petite révolution législative qui a inspiré de nombreuses autres initiatives à

commencer par le texte Californien, le CCPA (California Consumer Privacy Act), reconnu en 2024 dans une quinzaine d'États américains.

L'encadrement légal de la manipulation des données à caractère personnel est une brique clé de l'encadrement légal du développement et de l'utilisation de technologies d'IA. Le AI Act récemment voté au parlement européen en est la seconde brique toute aussi fondamentale, basée entre autres sur l'identification du risque associé à une IA, et qui pourra être mobilisé par le citoyen mais aussi et surtout le concepteur d'IA pour motiver sa décision vis-à-vis d'une technologie en particulier. La théorie des baïonnettes intelligentes devient alors un outil efficace en adéquation avec la loi qui la soutient, pour le concepteur face à son supérieur, au propriétaire de la solution qu'il développe, des actionnaires ou encore de ses utilisateurs ou clients.

La construction d'une gouvernance algorithmique (ou IA) d'excellence rend également possible l'usage de la théorie des baïonnettes intelligentes en fournissant à toutes les parties prenantes des bonnes pratiques et des lignes directrices leur permettant de légitimer leurs actions et leurs réactions. La création de standards comme ceux pour une IA dite de confiance tels que ISO/IEC 42001:2023, permettent également la constitution de ces lignes directrices. Une attention particulière doit être portée sur la nécessité d'encourager l'innovation par un cadre protecteur pertinent, tout en permettant l'expérimentation et la recherche à l'instar de ce qu'a réalisé le RGPD par la distinction explicite entre développement industriel et R&D (recherche et développement) dans sa mise en conformité.

***Conclusion***

Un tel cadre légal et de gouvernance qui met en œuvre la théorie des baïonnettes intelligentes, permettra également de construire une convention, un accord durable et explicite entre acteurs du marché autour de procédures tangibles [17]. Dans le cas des IAs dites de confiance, en évolution constante, il convient que ce cadre laisse la possibilité d'un ajustement continu des comportements individuels pour atteindre un équilibre stable. Cela s'accompagne aussi d'un partage libre des erreurs et des solutions envisagées et/ou testées, afin de contribuer collectivement aux apprentissages communs. La théorie des baïonnettes intelligentes trouve ses applications dans des domaines de plus en plus larges, la conception et l'usage d'intelligences artificielles en font partie pour défendre les droits fondamentaux individuels et collectifs, tout en luttant contre le risque de dilution des responsabilités.

***Références***


[1] *Algorithmes : la bombe à retardement* par Cathy O'Neil, publié chez *Les Arènes*, 2018 (initialement publié en anglais en 2016).

[2] *Contre-atlas de l'intelligence artificielle* par Kate Crawford, publié chez *Zulma*, 2022 (initialement publié en anglais en 2021)

[3] *How I'm fighting bias in algorithms*, par Joy Buolamwini, *TEDxBeaconStreet*, 2016.

[4] *Unmasking AI*, par Joy Buolamwini, publié chez *Random House*, 2023.

[5] *Les entreprises doivent construire une gouvernance algorithmique devançant la loi*, par Aurélie Jean, Guillaume Sibout, Mark Esposito et Terence Tse, *Harvard Business Review France*, 2023.

[6] *Explainable AI: A Brief History of the Concept*, par Mihály Héder, *ERCIM News*, numero 134 : pages 9 - 10, 2023.

[7] *Quand la machine apprend : La révolution des neurones artificiels et de l'apprentissage profond*, par Yann Le Cun, publié chez les éditions Odile Jacob, 2019.

[8] Revue générale du droit, *Conseil d'Etat, 10 novembre 1944, Sieur Langneur, rec. p. 248* : Revue générale du droit, numéro 10798, 1944.

[9] *Théorie du drone* par Grégoire Chamayou , publié chez les éditions La Fabrique, 2013.

[10] *The Political Philosophy of AI* par Mark Coeckelbergh, publié chez *Polity*, 2022.

[11] *La grande transformation* par Karl Polanyi, publié chez Gallimard, (1944 pour l'édition originale)

[12] *La trame de la négociation* par Anselm Strauss, publié chez L'Harmattan, 1992.

[13] *Propagations. Un nouveau paradigme pour les sciences sociales* par Dominique Boullier publié chez Armand Colin, 2023.

[14] *Technology and History: 'Kranzberg's Laws.'* par Melvin Kranzberg, publié dans Technology and Culture, vol. 27, no. 3, pp. 544–60, 1986.

[15] *Sociologie de la traduction. Textes fondateurs* par Akrich M, Callon M et Latour B publié aux Presses des Mines de Paris, 2006.

[16] *Cognition in the wild*, par Edwin Hutchins publié chez The MIT Press, 1995.

[17] *Analyse économique des conventions* par André Orléan (ed.) publié chez les Presses Universitaires de France, 2004.