LINC
CNIL.

**DOSSIER**

# Contrôle humain, décisions hybrides : quels enjeux ?

**Octobre 2024**
**linc.cnil.fr**

**Charlotte Barot,** Analyste IA

**Alors que les systèmes d'aide à la décision se répandent, leur usage dans des contextes décisionnels critiques soulève de sérieux problèmes éthiques et juridiques. Afin de prévenir les principaux risques identifiés dans le domaine de la prise de décision, la loi impose une intervention humaine ou un contrôle humain intégré dans la procédure de décision, aboutissant à des dispositifs « hybrides », combinant puissance de calcul et discernement humain. Dans cette série d'articles le LINC explore, d'après la littérature scientifique, deux obstacles à l'effectivité de tels dispositifs : d'une part les biais de confiance des utilisateurs vis-à-vis du système et, d'autre part, l'opacité des suggestions du système.**

# SOMMAIRE DU DOSSIER

# Introduction

Le contrôle humain se présente comme une mesure essentielle pour garantir des décisions fiables et justes, aboutissant à des procédures de décision hybrides, combinant intervention humaine et utilisation d'algorithmes. Cependant, ces procédures hybrides, qui visent à allier l'efficacité des systèmes d'aide à la décision et les qualités humaines de discernement, ne peuvent fonctionner que si le décisionnaire humain peut évaluer en toute connaissance de cause la sortie qui lui est proposée, comme la CNIL l'évoquait déjà en 2017 dans son rapport éthique « Comment permettre à l'Homme de garder la main ? »[1]. Dans cette revue de la littérature, le LINC explore les obstacles à la mise en œuvre de systèmes de décision hybride et les pistes identifiées pour leur amélioration.

# Accumulation des données : de nécessaires analyses

Face à l'essor des données massives, les algorithmes d'aide à la décision (ou ADM, « automated decision making ») se sont imposés pour traiter des problèmes complexes. Ces outils d'intelligence artificielle (IA) produisent des estimations rapides à partir desquelles les doivent prendre des décisions mais sans disposer d'éléments objectifs pour apprécier ce que ces IA proposent. On pense ici à des décisions dans des secteurs comme la santé (Jacobs 2021[2], Gaube *et al.* 2021[3], Beede *et al.* 2020[4]), la finance, la modération de contenu (Link *et al.* 2016[5], Gillespie 2020[6]), ou la détection de la fraude.

Les algorithmes d'aide à la décision présentent des bénéfices pour automatiser sur des tâches simples mais laborieuses ; ce faisant ils soulagent une partie de la charge de travail des humains et réduisent en principe les coûts en ressources humaines. A l'échelle individuelle, un ensemble de décisions mineures de la vie courante sont couramment déléguées à des algorithmes, sans conséquences dommageables, comme la recommandation du trajet le plus court pour rejoindre le lieu de travail ou une suggestion de morceaux à écouter.

Sur des tâches plus complexes avec des choix difficiles, la confiance dans ces outils peut aller jusqu'à leur déléguer non seulement l'exécution de tâches aux mécanismes bien connus, mais également une partie du jugement, afin d'en faire un guide, voire un « oracle ».

[1] https://www.cnil.fr/fr/comment-permettre-lhomme-de-garder-la-main-rapport-sur-les-enjeux-ethiques-des-algorithmes-et-de
[2] https://www.nature.com/articles/s41398-021-01224-x
[3] https://pubmed.ncbi.nlm.nih.gov/33608629/
[4] https://dl.acm.org/doi/10.1145/3313831.3376718
[5] https://idl.iscram.org/files/daniellink/2016/1401_DanielLink_etal2016.pdf
[6] https://www.researchgate.net/publication/343798653_Content_moderation_AI_and_the_question_of_scale

# Automatiser pour mieux décider

Si déléguer des décisions relevant du quotidien semble raisonnable, (car sans conséquences), demander à ChatGPT de prendre une décision importante, comme l'achat d'un bien immobilier, le semble moins. Un tel service est attendu pour fournir au mieux une recommandation ou un conseil, mais il semblerait déraisonnable, sur ce type de questions, de laisser à l'algorithme le dernier mot.

Ces systèmes, malgré leurs capacités impressionnantes, ne peuvent entièrement se substituer au jugement humain. Sur certains types de problèmes, ils s'avèrent inadaptés à des décisions qui requièrent de prendre en compte, en plus de la vraisemblance, un ensemble de facteurs relevant à la fois de l'éthique, du contexte social, et des considérations sur l'effet à long terme de la décision, comme en justice.

# Souveraineté humaine et décisions partagées

L'utilisation des systèmes automatisés soulève des problèmes éthiques, dont une réponse possible est l'intégration d'un contrôle humain pour préserver l'autonomie décisionnelle et garantir la pertinence de la décision.

Ainsi la loi « informatique et libertés » exclut-elle dans son article 47[7], les décisions de justice fondées sur un traitement automatisé et définit un cadre strict pour l'automatisation des décisions entraînant des conséquences sur les personnes qu'elles concernent. Elle demande que, lorsqu'une décision a des effets notables sur la vie d'un individu, celle-ci ne peut être l'issue d'une procédure entièrement automatisée, et doit ainsi inclure une intervention humaine, sauf dans le cadre de décisions administratives individuelles. Le Règlement Général sur la Protection des Données formule dans son article 22 les mêmes précautions et raffine le cadre des exceptions possibles : lorsque la personne concernée a donné son consentement, que le traitement automatisé est nécessaire à l'exécution d'un contrat, ou si une loi le prévoit explicitement.

# Efficacité des procédures

L'intervention humaine, malheureusement, n'améliore pas toujours les résultats produits par un système automatisé et peut même les dégrader. Cela s'explique par le fait que les décisionnaires peuvent adopter des attitudes rigides, soit en acceptant aveuglément les

[7] https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000037823131

suggestions (on parle de « *biais d'acceptation* »), soit en les rejetant par principe (on parle de biais d'aversion). Ces biais compromettent la performance et, dès lors, menacent la fiabilité des dispositifs hybrides.

- L'article **« Croire ou douter : Biais de confiance dans la prise de décision »** explore les questions relatives aux biais de confiance des décisionnaires humains, qui affectent la qualité des décisions et propose des pistes de remédiation identifiées par la communauté de recherche.

Les problèmes de confiance ne sont cependant pas de simples erreurs de jugement : ils révèlent une difficulté plus profonde à évaluer correctement les résultats fournis par l'IA, soulevant la question de l'interprétabilité des sorties. Il est donc crucial de doter les décideurs des outils et des connaissances nécessaires pour déterminer quand suivre les recommandations des algorithmes, afin d'optimiser l'efficacité du dispositif global, et de répartir la charge de la décision.

- L'article **« Prédire sans expliquer, ou quand l'opacité algorithmique brouille les cartes** » explore les enjeux d'intelligibilité des résultats des systèmes automatisés, qui peuvent être difficiles à interpréter ou à remettre en question et détaille des modalités pratiques pour faciliter l'interaction avec les sorties proposées et renforcer la fiabilité des procédures de prise de décision hybride.

# Croire ou douter : la question des biais de confiance dans la prise de décision

En décembre 2023, la Cour de Justice de l'Union Européenne[8] a statué que l'outil de *credit scoring* de la société allemande SCHUFA, qui produisait pour les banques une estimation de solvabilité d'un client sous la forme d'un score de confiance, constitue une décision entièrement automatisée. Pourtant, la décision d'octroi de crédit était effectivement prise par un employé de la banque chargé de vérifier et appliquer ou non la proposition de l'outil, ce qui impliquait une intervention humaine. Dans la mesure, cependant, où ce score n'était jamais contesté par les employés de la banque qui s'y fiaient systématiquement, transformant ainsi la suggestion en décision d'attribution de crédit, la Cour a estimé que cette intervention n'était pas significative.

Pour ne pas être qu'une simple formalité, le contrôle humain, tel que défini dans le Règlement européen sur l'IA[9], doit permettre de détecter des erreurs, de diverger ou d'interrompre le système, autrement dit de fournir une véritable alternative à la sortie du système d'IA. Or, ces compétences reposent sur des dispositions psychologiques que le règlement souligne, en indiquant que la personne chargée du contrôle doit avoir conscience des biais cognitifs possibles, comme une excessive confiance. Ce biais constitue un élément risquant d'altérer la qualité du contrôle exercé, et donc la fiabilité du dispositif dans son ensemble.

La littérature empirique ne converge pas vers une réaction uniforme des individus aux algorithmes mais identifie deux tendances : soit une heuristique d'acceptation (thèse de l'appréciation) identifiée par le Règlement européen sur l'IA, soit de rejet (thèse de l'aversion), en dépit des erreurs que ces stratégies génèrent.

# Trop de confiance attire le danger : appréciation excessive

Certaines études pointent une tendance des participants à se conformer aux sorties du système (e.g. Jacobs et al. 2021[10], Green 2019[11], Yin 2019[12], Bussone et al. 2015[13], Kiani et al.

[8] https://curia.europa.eu/juris/document/document.jsf;jsessionid=E3D1DC708C777FB310636E7197610C45?text=&docid=280426&pageIndex=0&doclang=fr&mode=lst&dir=&occ=first&part=1&cid=4793519

[9] https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[10] https://www.nature.com/articles/s41398-021-01224-x

[11] https://scholar.harvard.edu/bgreen/publications/disparate-interactions-algorithm-loop-analysis-fairness-risk-assessments

[12] https://dl.acm.org/doi/10.1145/3290605.3300509

[13] https://ieeexplore.ieee.org/document/7349687

2020[14], Alberdi et al. 2004[15], Logg et al. 2019[16], Robinette et al. 2017[17]). Ainsi, le contrôle humain ne serait plus effectif car le décisionnaire se fierait systématiquement à la sortie du système, dans la même logique que celle de l'arrêt de la Cour de Justice Européenne concernant SCHUFA. Dans ce cas, ce n'est pas l'efficacité de la décision qui est remise en question mais l'effectivité du contrôle humain sur la décision algorithmique, et sa capacité à rejeter les éventuelles erreurs du système ou anomalies.

Dans une étude menée en psychiatrie (Jacobs et al. 2021[18]), des chercheurs demandent à une cohorte de cliniciens de prendre une série de décisions sur un patient fictif. Pour chaque patient, le clinicien doit décider d'un traitement, soit avec une recommandation d'un système d'apprentissage automatique et une brève justification de la réponse choisie, soit sans aucune recommandation (choix complètement indépendant).

L'étude montre que, d'une part, les performances des groupes ayant et n'ayant pas accès aux recommandations du système d'apprentissage automatique sont virtuellement les mêmes, et que les deux groupes prennent de moins bonnes décisions que le système seul. D'autre part, lorsque le système produit une décision erronée, (où l'erreur est définie comme une position divergeant avec celle d'un collège d'experts en psychiatrie), la performance baisse par rapport au groupe de contrôle (sans accès au système d'apprentissage automatique), c'est-à-dire que l'humain tend à se conformer à la décision de l'algorithme sur ce type de cas.

Enfin, les auteurs constatent un effet de la familiarité avec l'outil : les cliniciens plus familiers avec le système étaient moins susceptibles en moyenne d'utiliser une recommandation du système d'apprentissage automatique, quelle que soit son exactitude, par rapport aux cliniciens moins familiarisés avec les systèmes d'apprentissage automatique.

Le constat selon lequel le niveau d'expertise « métier » joue un rôle dans la prise de décision est corroboré par une étude de Gaube et al. 2021[19] dans laquelle deux groupes de médecins de niveau d'expertise médicale différent doivent produire un diagnostic et noter une recommandation affichée comme provenant soit d'un algorithme soit d'un humain, alors que toutes les recommandations proviennent d'un humain.

Les médecins du groupe le plus expert tendent à moins bien noter les recommandations lorsqu'elles sont indiquées comme provenant d'un système d'IA. En revanche, la qualité de leur diagnostic est influencée par la qualité du conseil reçu, indépendamment de sa provenance (IA ou humain).

Ce résultat suggère que les effets d'influence du conseil du système pourraient être un simple artefact du fait de n'avoir pas pu former sa décision avant la confrontation, et pas nécessairement due à une attitude de déférence envers l'algorithme. En effet, l'étude montre

[14] https://www.nature.com/articles/s41746-020-0232-8
[15] https://pubmed.ncbi.nlm.nih.gov/15354301/
[16] https://www.sciencedirect.com/science/article/abs/pii/S0749597818303388
[17] https://ieeexplore.ieee.org/document/7828078
[18] https://www.nature.com/articles/s41398-021-01224-x
[19] https://www.nature.com/articles/s41746-021-00385-9

que les participants ont globalement eu tendance à suivre les conseils prodigués, qu'ils proviennent d'un humain ou d'une machine.

L'absence de déférence envers le système par des experts est aussi constatée dans d'autres études (Logg et al. 2020[20], Povyakalo et al. 2013[21]). Dans cette dernière étude, l'appréciation était aussi modulée par le désaccord. Un conseil en contradiction avec l'opinion préalable du participant avait moins d'impact mais n'annulait pas complètement l'effet de la confiance envers le système. L'appréciation des algorithmes a diminué (mais n'a pas disparu) lorsque leurs conseils ont été opposés à leur propre jugement. C'est donc sur ces points de désaccord, en concluent les auteurs, que les gens sont les plus susceptibles d'améliorer leur précision. Ces cas critiques de confrontation sont donc intéressants pour la prise de décision.

## Aversion et méfiance excessive

Certaines études suggèrent, elles, une méfiance excessive envers le système, empêchant les participants de prendre des décisions éclairées (e.g. Dietvorst et al. 2015[22], Longoni et al. 2019[23], Dzindolet et al. 2002[24], Lim et O'Connor 1995[25], Yeomans et al. 2019[26], Promberger et Baron 2006[27]).

Dans plusieurs tâches de prévision où ils doivent choisir entre une prédiction algorithmique et une prédiction humaine concernant la vraisemblance du succès d'un titre musical (Dietvorst et al. 2015[28]), lorsqu'ils ont vu le système fonctionner, et parfois se tromper, les participants tendent à rejeter ses prédictions au profit de conseils humains, et ce en dépit du taux d'erreur plus élevé des prédictions humaines (allant jusqu'à doubler par rapport à celles de l'algorithme).

En somme, les humains pardonnent moins facilement leurs erreurs aux algorithmes, et ont tendance à généraliser l'ensemble de leur performance sur la base de ces exemples délétères. Cette tendance persiste même après avoir observé que les performances du système dépassent en moyenne celles des humains. Une explication naturelle est qu'une méfiance envers les algorithmes existe avant même de les avoir observés, et qu'observer des erreurs renforce cet *a priori*.

Il y aurait donc à l'œuvre un biais d'ancrage : une fois leur opinion à propos d'un système d'IA arrêtée, il est très difficile pour les utilisateurs de changer d'avis, même en présence de preuves contradictoires. Ceci est soutenu par le fait que l'aversion aux algorithmes n'est pas

[20] https://www.sciencedirect.com/science/article/abs/pii/S0749597818303388
[21] https://pubmed.ncbi.nlm.nih.gov/23300205/
[22] https://psycnet.apa.org/record/2014-48748-001
[23] https://www.researchgate.net/publication/332550851_Resistance_to_Medical_Artificial_Intelligence
[24] https://journals.sagepub.com/doi/abs/10.1518/0018720024494856
[25] https://psycnet.apa.org/record/1996-10296-001
[26] https://onlinelibrary.wiley.com/doi/abs/10.1002/bdm.2118
[27] https://psycnet.apa.org/record/2006-23369-005
[28] https://psycnet.apa.org/record/2014-48748-001

observée dans des tâches purement déterministes comme des calculs logiques, ou des tâches de mémoire, sur lesquelles les humains sont notoirement plus défaillants que les algorithmes.

# Amplification des biais et interactions délétères

Les humains ne sont pas dépourvus de biais et n'ont pas un jugement infaillible, ce qui induit des erreurs s'ajoutant, toutes choses égales par ailleurs à celles des systèmes, ou les amplifiant.

**LE CAS DE COMPAS**

Le système **COMPAS** (Correctional Offender Management Profiling for Alternative Sanctions) est un outil de prédiction utilisé dans le système de justice pénale aux États-Unis. Développé par la société Equivant (ex Northpointe), ce système évalue le risque qu'un délinquant commette de nouvelles infractions ou ne se présente pas à une audience future. COMPAS utilise un ensemble de questions sur les antécédents criminels, le comportement, et les caractéristiques personnelles des délinquants (relations sociales, données démographiques : âge, sexe, origine ethnique, etc.) pour calculer des scores de risque, dont celui de récidive violente. Les juges utilisent souvent les scores COMPAS pour décider si un prévenu peut être libéré en attente de son procès. Une enquête menée par le média **ProPublica** en 2016, a révélé que COMPAS présentait des biais discriminatoires, surestimant le risque de récidive chez certains individus et induisant des peines plus lourdes, sans que ceci soit justifié par d'autres critères que l'origine ethnique.

Pour comprendre si ces biais sont uniquement hérités de l'algorithme, **Green** et al. 2019 reproduisent le cas du système COMPAS en conditions de laboratoire. La tâche expérimentale consistait à évaluer le risque de récidive d'un individu, dans la même logique que l'algorithme original. La description de l'individu était assortie d'un score de risque de récidive en pourcentage, et le participant devait indiquer un score de récidive. Les chercheurs constatent qu'en présence d'une suggestion de l'algorithme, les individus prennent une moins bonne décision que ce dernier, et sont incapables d'évaluer tant leur performance que celle de l'algorithme. Ils soulignent que les décisionnaires sont manifestement eux-mêmes biaisés en défaveur de certains profils, ce qui les conduit à amplifier le biais de l'algorithme. Les biais humains, combinés à une haute confiance en l'algorithme, peuvent donc conduire à accepter le résultat biaisé voire à amplifier sa tendance.

A ce sujet voir également les interviews d'**Angèle Christin** *(«Les méthodes ethnographiques nuancent l'idée d'une justice prédictive et entièrement automatisée »*)et de **Philippe Besse** ( *« Les décisions algorithmiques ne sont pas plus objectives que les décisions humaines »*) sur le site du LINC.

# Jugement adéquat : influences et conditions d'exercice

Les constats rappelés ci-dessus incitent à porter une attention particulière aux conditions permettant à l'utilisateur d'être dans les meilleures dispositions psychologiques possibles, afin d'exercer un jugement pertinent. Ceci concerne à la fois ses compétences propres, la tâche à résoudre, mais aussi un ensemble de facteurs exogènes liées à l'environnement de travail et au contexte de la prise de décision. On peut ainsi relever :

L'environnement de la prise de décision :

- ***Le coût de l'erreur ou le fait d'être engagé dans une situation risquée*** (e.g. Robinette et al. 2017[29]). Les risques pris par la personne qui prend la décision compte tenu des conséquences qu'elle entraîne pour la personne concernée influencent son jugement, car suivre ou diverger de la proposition d'un système peut entrainer des conséquences différentes en cas d'erreur. En particulier, ***l'attribution de responsabilité au preneur de décision,*** qui est un coût pesant sur le décisionnaire. Si la responsabilité d'une décision est attribuée à l'algorithme, il pourra sembler coûteux pour l'humain qui intervient de diverger avec la décision de l'algorithme, ce qui l'incite à accepter les décisions de l'algorithme
- ***Le temps alloué à la prise de décision*** (e.g. Robinette et al. 2017[30]). Une durée très courte de délibération entraine des prises de décisions se rapprochant des suggestions de l'algorithme, par un réflexe d'économie.

L'expertise métier :

- ***Le niveau d'expertise dans la tâche donnée*** (e.g. Jacobs et al.. 2021[31], Logg et al. 2019[32], Povyakalo et al.. 2013[33]). Les experts tendent à moins se fier à l'algorithme qu'une personne plus novice, ce qui est un avantage pour la possibilité de diverger de la décision proposée par l'algorithme, mais peut également pousser l'expert dans une voie de confiance peut être excessive. A l'inverse les personnes plus novices risquent d'accepter la suggestion du système plus souvent.
- ***Le doute vis-à-vis de la décision à prendre***. Les cas où la personne en charge de la décision n'est pas certaine ou a un niveau de doute assez élevé seront vraisemblablement ceux où l'influence de la suggestion sera la plus grande. Il peut s'agir de cas difficiles car les options sont trop nombreuses, ou parce que le cas rencontré est singulier.

[29] https://ieeexplore.ieee.org/document/7828078
[30] https://ieeexplore.ieee.org/document/7828078
[31] https://www.nature.com/articles/s41398-021-01224-x
[32] https://www.sciencedirect.com/science/article/abs/pii/S0749597818303388
[33] https://pubmed.ncbi.nlm.nih.gov/23300205/

La relation aux systèmes automatisés :

- ***La familiarité avec les algorithmes et les systèmes automatisés*** (e.g. Jacobs et al. 2021[34]). Les personnes moins familières avec ces systèmes ont tendance, selon les cas, à avoir plus confiance, ou au contraire une méfiance excessive, mais la disposition est rarement neutre.
- ***La confiance a priori envers le système***, i.e. les préjugés sur la fiabilité de l'algorithme, qui s'avèrent très difficiles à déconstruire, en particulier si ceux-ci sont fondés sur l'observation que le système a fait des erreurs (e.g. Robinette et al. 2017[35], Dietvorst et al. 2015[36], Prahl et al. 2017[37]).
- ***La congruence, ou le fait d'être d'accord intuitivement avec la sortie de l'algorithme*** (e.g. Logg et al. 2019[38]). Ce facteur, en apparence trivial, risque d'avoir un fort impact sur la capacité humaine à se remettre en question dans des cas de méfiance excessive envers l'algorithme.

Enfin, la configuration globale dans laquelle s'exerce l'interaction entre le décisionnaire et le système d'aide à la décision :

- ***La suggestion anticipée.*** Le système propose une recommandation avant même la prise de décision humaine. L'humain décide s'il choisit d'accepter la décision ou s'il choisit de diverger et d'en proposer une autre. Comme vu précédemment, cette configuration risque de favoriser les effets d'ancrage et ainsi de rendre plus difficile pour l'humain de diverger.
- ***La levée de doute.*** Le système ne déclenche aucune intervention, mais signale une situation sur laquelle il identifie un danger. Dans ce dernier cas seulement, il y a intervention humaine, quand un employé décide de valider ou non l'alerte. Souvent utilisée dans la télésurveillance, cette configuration permet d'économiser l'implication humaine car le personnel n'intervient que sur certains types de sorties : celles indiquant un danger ou la nécessité d'une intervention. Il nécessite que les situations potentiellement dangereuses soient bien toutes identifiées par le système.
- ***La suggestion alternative.*** Le système produit une information supplémentaire en support de la décision humaine déjà envisagée. Cette solution permet d'éviter l'excès de confiance en l'algorithme, mais peut en revanche, à affaiblir l'efficacité de son utilisation de, car l'opérateur humain a déjà pris sa décision et peut trouver « coûteux » d'en changer. En revanche, sur les cas où l'humain a un doute quant à sa décision, la configuration semble particulièrement pertinente.
- ***L'utilisation d'un algorithme de choix.*** Un algorithme de choix décide de renvoyer la prise de décision à l'humain ou au système seul (décision unique, non contestée). Cette solution proposée par Mozannar et al.[39] 2023 permet d'éviter les biais de confiance

[34] https://www.nature.com/articles/s41398-021-01224-x
[35] https://ieeexplore.ieee.org/document/7828078
[36] https://psycnet.apa.org/record/2014-48748-001
[37] https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2464
[38] https://www.sciencedirect.com/science/article/abs/pii/S0749597818303388
[39] https://proceedings.mlr.press/v206/mozannar23a/mozannar23a.pdf

humains, mais il repose sur un autre système automatisé, ce qui ne satisfait pas les exigences du RGPD et du Règlement sur l'IA, sauf à démultiplier les niveaux d'intervention humaine.

- ***L'indice dans un faisceau.*** Indépendamment, un humain et une IA prennent une décision sur un problème ; un tiers, donc un collège ou un autre expert, tranche sur les désaccords, idéalement à l'aveugle.

Les options de « **confrontation** », donc les options 3 et 5, dans lesquelles l'utilisateur a la possibilité de se faire une opinion autonome, semblent les plus à mêmes de permettre à la personne de ne pas « subir » les décisions du système, et donc d'exercer son pouvoir de jugement. En revanche, pour qu'elles soient efficaces, il faut que l'humain soit enclin à véritablement confronter sa décision avec la suggestion du système reçue ensuite, sans nécessairement tenter de confirmer sa propre hypothèse. Si ces deux configurations semblent ainsi pertinentes dans le cadre du contrôle humain, il est probable que les cas véritablement profitables seront ceux sur lesquels l'avis de l'agent n'est pas trop tranché, au risque sinon de rejeter la décision.

# Conclusion

Même avec une neutralité parfaite, un jugement infaillible, et des conditions de travail idéales, il reste toutefois un obstacle fondamental : la possibilité même de déchiffrer et de donner un sens à la suggestion du système. C'est cette difficulté intrinsèque que nous proposons d'explorer dans l'article « Prédire sans expliquer, ou quand l'opacité algorithmique brouille les cartes ».

# Prédire sans expliquer, ou quand l'opacité algorithmique brouille les cartes

Le second article de cette série illustrait les biais de confiance à l'œuvre dans un contexte de prise de décision assistée, et en quoi ils constituent un obstacle au libre exercice du jugement humain dans un dispositif hybride. Si ces biais reposent sur des aptitudes particulières et des éléments de contexte propres à la prise de décision, ils sont aussi symptomatiques d'une difficulté intrinsèque au fonctionnement du système : la possibilité d'interpréter les sorties. En effet, lorsque la personne en charge de prendre la décision ne parvient pas à évaluer la suggestion du système, elle ne peut qu'avoir recours à sa propre intuition ou décider de suivre une heuristique de remise en doute ou de confiance. Même la condition minimale exigible d'un contrôle humain, qui serait celle d'écarter les cas évidents d'erreur, n'est pas toujours réaliste dans un contexte où certaines sorties sont complexes à évaluer. D'une part, des réponses absurdes se présentent parfois comme des faits vraisemblables : c'est le cas des systèmes d'IA génératives de texte qui peuvent insérer dans un texte correct un nom ou un fait inventé. D'autre part, le format même de la sortie peut le rendre étanche à l'évaluation : juger du bien-fondé d'un score numérique pour proposer une alternative supposerait dans la plupart des cas de refaire l'inférence qui a abouti à ce score.

## Prédire n'est pas expliquer

En 1999, dans une étude comportementale, Goodwin et Fildes[40] établissent que, lorsqu'on leur soumet des prédictions de tendances dans le domaine du marketing, les décisionnaires tendent à, au mieux, ignorer des prédictions fiables, voire à les dégrader en tentant de les modifier. Ils ont donc tendance à montrer un biais de méfiance envers l'algorithme. Or, les auteurs relèvent que les sorties sont difficiles à évaluer par les décisionnaires car leur format, donné sous la forme d'un score ou d'un pourcentage, n'est pas facilement contestable. Si les utilisateurs montrent de telles attitudes inadaptées, c'est parce qu'ils ne sont pas capables de déchiffrer le score qui leur est donné, et qu'ils préfèrent donc l'ignorer la plupart du temps. Lorsqu'ils tentent, cependant, de proposer une alternative, ils ne parviennent pas à faire mieux que le système.

Comme le montre cet article, les personnes chargées d'évaluer les sorties sont finalement chargées de deux sous-tâches : comprendre et évaluer. D'une part, l'utilisateur doit donner du sens à la suggestion produite : par exemple, si cette sortie est un score, comprendre sur quelle échelle figure ce score et quels sont les seuils considérés comme critiques. L'utilisateur

[40] https://onlinelibrary.wiley.com/doi/epdf/10.1002/%28SICI%291099-0771%28199903%2912%3A1%3C37%3A%3AAID-BDM319%3E3.0.CO%3B2-8

doit pouvoir lire le message global envoyé par le système (chiffre), dans son contexte (ici l'échelle, et les seuils).

Ensuite, l'utilisateur doit l'évaluer, c'est-à-dire produire un jugement sur sa pertinence, pour l'accepter, ou la rejeter au profit de sa propre opinion ou d'une version corrigée. Dans des contextes d'utilisation de systèmes d'apprentissage machine, évaluer les sorties n'est pas trivial car ces modèles opèrent en boîte noire. Lorsque l'on ne peut pas retracer et décortiquer l'inférence qui a produit la suggestion du système, il faut tout de même trouver des moyens d'interpréter les sorties.

# Vers des systèmes introspectifs

Une option pourrait être de poser la question au système lui-même, afin qu'il produise des justifications permettant à la fois de comprendre sa sortie et d'évaluer sa propre fiabilité. Malheureusement, les justifications et les scores de confiance accompagnant les réponses ne sont pas toujours fiables, même lorsque les sorties sont correctes.

Ainsi Jin *et al.* 2024[41], analysant les performances d'un modèle chargé de résoudre des cas cliniques à partir de lecture d'images, ont-ils constaté les faibles compétences du modèle pour justifier ses réponses. Le modèle a été testé en utilisant des prompts structurés en trois parties : il devait d'abord décrire l'image médicale fournie, rappeler des informations médicales pertinentes pour répondre à la question posée, produire un raisonnement médical et enfin choisir un diagnostic parmi un ensemble d'options. Si le modèle montrait une grande précision -parfois supérieure à celle des médecins- dans ses diagnostics finaux, il était mis à rude épreuve dans sa compréhension des images médicales, ce qui le poussait à fournir des raisonnements bancals pour appuyer un diagnostic pourtant correct, produisant ainsi des justifications trompeuses.

Ces limites rendent son utilisation en milieu clinique encore prématurée, car ces faiblesses menacent une potentielle intégration dans la pratique médicale. Le risque d'introduire des justifications trompeuses, par exemple dans un cas où l'expert humain n'aurait pas accès à l'image, ou se reposerait sur la fausse lecture de l'image produite, serait d'induire en erreur la personne prenant la décision, pouvant même conduire à rejeter la suggestion finale correcte.

# Analyse comportementale des systèmes

En somme les systèmes n'ont pas toujours de bonnes capacités d'introspection : ils ne sont pas toujours capables d'analyser eux-mêmes leur propre comportement, qu'il soit bon ou

[41] https://www.nature.com/articles/s41746-024-01185-7

mauvais. En revanche, on peut toujours se guider sans tenter d'ouvrir la boîte noire : à l'aide d'une analyse comportementale du modèle.

C'est à ce titre qu'intervient le développeur d'un système dans la bonne intégration du modèle dans un processus d'expertise métier. Il peut fournir plusieurs éléments de contexte donnant une idée plus précise des conditions dans lesquelles le modèle a été « élevé », aidant à interpréter son comportement :

- sur le contexte dans lequel l'algorithme a été conçu,
- sur ses limites connues,
- sur des tests effectués avant la mise sur le marché,
- sur des tâches sur lesquelles il se montre typiquement moins performant, etc.

Les informations sur les données d'entraînement, le comportement en situation et les marges d'erreurs associés aux tests, permettent d'éclairer également les sorties.

# Apprentissage par renforcement des utilisateurs

Des recherches exploratoires développent en profondeur cette notion d'analyse comportementale des modèles en proposant aux utilisateurs des modèles « d'entraînement » des décideurs humains pour les familiariser avec le comportement du système utilisé (Lian et Tan 2019[42], Suresh *et al.*. 2021[43]; Wortman Vaughan et Wallach 2021[44]). L'objectif est d'apprendre aux utilisateurs par un certain nombre d'essais à se familiariser avec le comportement du système et ainsi de savoir détecter quand suivre ses suggestions, ou quand les rejeter, et auquel cas creuser le problème en profondeur.

Dans leur dispositif expérimental, Mozannar *et al.* 2022[45] explorent l'optimisation de la collaboration entre humains et systèmes d'intelligence artificielle sur des tâches de réponses à des questions basées sur des passages de textes (basé sur le jeu de données HotPotQA[46]). L'article propose une méthode pour aider les utilisateurs à collaborer avec différents modèles d'IA : au terme de l'entraînement, ils doivent parvenir à décider quand il est préférable de déléguer la réponse au modèle, et quand ils devraient intervenir.

Cette méthode s'inspire de recherches en éducation soulignant l'importance du retour sur expérience dans l'apprentissage. Elle se base sur le principe des exemples spécifiques, qui sont des cas-type destinés à illustrer les situations où l'algorithme est fiable et celles où il ne l'est pas. Les exemples sont choisis pour représenter différents scénarios : certains où l'algorithme présente un haut niveau de confiance et est correct dans sa prédiction, d'autres où le niveau

[42] https://arxiv.org/abs/1811.07901
[43] https://dl.acm.org/doi/pdf/10.1145/3490099.3511160
[44] https://www.jennwv.com/papers/intel-chapter.pdf
[45] https://arxiv.org/abs/2111.11297
[46] https://hotpotqa.github.io/

de confiance est élevé mais la prédiction fausse, ainsi que des cas où le niveau de confiance est incertain, que la prédiction soit correcte ou non.

L'objectif est d'améliorer le « modèle mental » que les humains se font des capacités de l'algorithme, c'est-à-dire de les amener à comprendre les cas sur lesquels il est susceptible de faire des erreurs, y compris sur ses propres estimations de confiance. Ce processus d'apprentissage permet aux utilisateurs de mieux comprendre les situations dans lesquelles ils peuvent faire confiance à l'algorithme et celles où, au contraire, il est nécessaire de vérifier les résultats plus attentivement.

Les expériences montrent que les utilisateurs formés avec cette méthode sont plus efficaces pour décider quand déléguer les décisions au classificateur, améliorant la collaboration entre systèmes de décisions et humains et améliorant les erreurs de jugement.

# Conclusion générale

La littérature scientifique montre que la mise en œuvre de systèmes de décision hybrides présente deux types d'enjeux : d'abord permettre au décisionnaire d'exercer un jugement en principe éclairé et impartial, ce qui relève de conditions exogènes de prise de décision, et ensuite permettre au décisionnaire de lire correctement les sorties, ce qui relève de conditions intrinsèques de lisibilité du système. Au fond, les attitudes de confiance du décisionnaire ne sont que le reflet de ces conditions initiales dont il hérite. Finalement, ces deux types d'obstacles sont donc signe que la responsabilité de la décision hybride est une charge à répartir entre le déployeur du système, qui a en charge le risques métier, et le concepteur du système, qui doit répondre du bon fonctionnement de son système et fournir un certain nombre de clefs pour apprendre à l'utiliser.

En effet, si l'intervention humaine requiert une marge de manœuvre importante du décisionnaire, le risque est d'augmenter, par symétrie, sa responsabilité individuelle dans la prise de décision : plus sa liberté est grande, plus on fait peser sur lui les coûts associés. Au-delà de la procédure de décision en elle-même, il faut donc élargir l'échelle et penser ces nouvelles procédures dans l'ensemble du contexte de travail pour intégrer de la meilleure manière les suggestions des machines aux décisions des humains.