Exiger l'explicabilité de l'IA : Une coquetterie humaine.

9 mars 2026

Et elle nous coûte des vies.

Nous exigeons des algorithmes médicaux qu'ils « s'expliquent » (ou à tout le moins, que nous soyons capables de comprendre et d'expliquer leurs « décisions »).

C'est devenu un impératif réglementaire, éthique, presque moral. L'AI Act le grave dans le marbre. Les comités d'éthique le psalmodient. Les conférences académiques lui consacrent des tracks entiers.

Les start-ups d'IA en santé s'en gargarisent à l'envi. « Notre IA est explicable : elle repose sur de l'IA symbolique et des réseaux bayésiens. » Sous-entendu : elle saurait ce qu'elle fait. Mieux que nous.

Traçable, oui (au prix d'un surcoût computationnel que personne ne chiffre dans les démos). Compréhensible, peut-être.

Une seule question : Vous pouvez m'expliquer votre dernière décision importante ?

Pas la version que vous en donnerez, une élaboration construite après coup, lissée, cohérente, présentable. C'est à dire aussi solide qu'une introspection psychanalytique...

Jérôme Vetillard

CTO | VP R&D | Chief Product Officer | AI-Powered Healthcare & Life Sciences Products | Compliance by Design | PhD AgroParisTech | CPO MIT Sloan | Exec MBA IE Business School & Brown University

Twingital-institute / Twingital-ventures : twingital-ventures.com

Je veux le vrai processus. L'activation réelle de vos circuits de neurones. La séquence causale depuis le stimulus jusqu'à l'acte. La raison pour laquelle ce souvenir plutôt qu'un autre a pesé. Pourquoi cette heuristique a dominé celle-là.

Vous ne pouvez pas. Personne ne peut.

Voici une expérience de pensée. Imaginez un IRM hypothétique (IRM 1000 Tesla) capable d'imager en temps réel l'activité de chaque neurone, chaque synapse, chaque gradient électrochimique de votre cerveau au moment où vous prenez une décision clinique difficile.

Auriez-vous expliqué cette décision ? Non. Vous auriez une carte. Exhaustive, précise, spectaculaire. Mais une carte n'est pas une loi. Une description n'est pas une explication.

Savoir ce qui s'est passé physiologiquement dans les réseaux de neurones, n'est pas savoir pourquoi ces activations produisent ce choix et pas un autre. C'est la distinction entre syntaxe et sémantique. Entre le comment et le pourquoi. Entre l'observation et la compréhension.

Nous n'avons pas cette explication pour le cerveau humain. Nous ne l'aurons probablement jamais sous cette forme.

Les ingénieurs IA, "sommés" de trouver une réponse technologique, répondent par la traçabilité, comme si documenter le chemin équivalait à comprendre pourquoi ce chemin. Traçabilité et explicabilité ne sont pas synonymes.

Et cependant, nous l'exigeons de l'algorithme. Il doit être "explicable".

Mais explicable… pour qui ?

Le mot « explicable » est un prédicat relationnel. Il n'existe pas dans l'absolu !

Il existe pour quelqu'un, selon un modèle mental, dans un contexte donné.

Explicable pour un radiologue senior ? Un interne ? Un patient de 72 ans ? Un juge administratif ? Un régulateur de l'ANSM ? Un expert IA ? Ces six personnes n'ont pas le même espace conceptuel pour recevoir une explication.

Toute explication est une projection : un passage vers un espace de dimension inférieure. Elle est calibrée sur le destinataire, jamais sur le processus réel."

Ce que nous appelons « explication » en XAI - eXplainable AI- (LIME, SHAP, cartes d'attention, contrefactuels) produit des approximations locales d'un comportement global. C'est honnête sur le plan technique.

Jérôme Vetillard

CTO | VP R&D | Chief Product Officer | AI-Powered Healthcare & Life Sciences Products | Compliance by Design | PhD AgroParisTech | CPO MIT Sloan | Exec MBA IE Business School & Brown University

Twingital-institute / Twingital-ventures : twingital-ventures.com

Mais personne ne dit à voix haute la chose embarrassante : ces approximations peuvent être entièrement fausses sur le plan mécanistique, et quand même augmenter la confiance de l'utilisateur. Des études le démontrent. Des explications incorrectes génèrent de la confiance. L'explication remplit une fonction rhétorique, pas épistémique.

Un bastion coquet de l'anthropocentrisme

Ce double standard n'est pas une erreur scientifique. **C'est une posture anthropocentrique défensive.**

On impose à la machine un standard de justification parce qu'on est profondément mal à l'aise avec une idée simple et dérangeante : une entité « non-consciente » peut prendre de meilleures décisions que nous sans (s)avoir (à) s'en expliquer. Et nous, nous n'avons jamais eu à le faire.

Un cardiologue expert qui rate 25% des NSTEMI sur ECG (infarctus dont la présentation électrocardiographique est souvent invisible à l'œil humain entraîné) en racontant une belle histoire clinique est objectivement moins fiable qu'un algorithme de deep learning atteignant des performances comparables ou supérieures sur la même tâche, sans être en capacité architecturale à s'expliquer.

La méfiance envers la décision algorithmique a des décennies. L'algorithme capable de surpasser le clinicien n'en a que cinq. Nous avons choisi de faire confiance au cardiologue avant même que la question de la fiabilité du diagnostic ne puisse se poser.

Non par rationalité. Par confort ontologique. Par coquetterie.

Sous-entendu collectif, rarement formulé : nous, humains, serions capables d'expliquer rationnellement et sans biais nos décisions. L'algorithme, non. C'est précisément ce présupposé que cinq décennies de neurosciences cognitives ont méthodiquement démoli.

Ce qui devrait être vraiment exigible

Arrêtons de nous tirer une balle dans le pied réglementaire.

Si l'objectif réel est la sécurité des patients, la question n'est pas :

- « pourquoi l'algorithme a-t-il décidé ça » ?

c'est :

- « dans quelle mesure peut-on s'y fier, et quand peut-il se tromper » ?

Jérôme Vetillard

CTO | VP R&D | Chief Product Officer | AI-Powered Healthcare & Life Sciences Products | Compliance by Design | PhD AgroParisTech | CPO MIT Sloan | Exec MBA IE Business School & Brown University

Twingital-institute / Twingital-ventures : twingital-ventures.com

Ce sont des questions mesurables. Auditables. Actionnables. Ce sont des questions de fiabilité. En termes opérationnels :

- Variance inter-runs sur input identique. Si vous soumettez le même ECG deux fois à 3 minutes d'intervalle et obtenez des résultats différents, vous avez un problème et vous n'avez pas besoin d'une explication, vous avez besoin d'une correction.
- Calibration de l'incertitude. Le modèle sait-il quand il ne sait pas ? Un système qui produit « probabilité 94% » sur un cas qu'il n'a jamais vu est dangereux. Un système qui dit « je suis en dehors de mon domaine de validité » est utile.
- Stabilité face aux perturbations non-significatives. Une variation de ±2 mmHg sur la tension systolique ne devrait pas retourner un diagnostic. Si c'est le cas, le modèle est fragile, qu'il puisse s'expliquer ou non.
- Stratification des erreurs. Pas un taux d'erreur global, une cartographie des erreurs. Où le modèle échoue-t-il ? Sur quelles populations ? Dans quelles conditions cliniques ? C'est actionnable.
- Détection de dérive. Le modèle se dégrade-t-il silencieusement sur de nouvelles données ? C'est une question de surveillance post-commercialisation, exactement comme pour un médicament : Des concepts de pharmacovigilance appliqués aux dispositifs médicaux IA.

Conclusion

La vraie transparence n'est pas mécanique. Elle est systémique.

Elle ne demande pas à l'algorithme d'expliquer chacune de ses décisions. Elle demande aux développeurs de documenter les données d'entraînement, les limites connues, les populations de validation. Elle demande aux intégrateurs de vérifier l'adéquation au contexte clinique. Elle demande aux cliniciens d'assumer la décision finale, informés par le système.

C'est un modèle de responsabilité distribuée (le même qui prévaut dans l'aviation, le nucléaire, la pharmacologie). Non pas parce que ces systèmes sont architecturalement comparables à un algorithme de deep learning : un pilote automatique est déterministe, formellement vérifiable, certifiable par preuve mathématique. Mais parce que le modèle de gouvernance est transposable : certifier le système, auditer ses performances, surveiller sa dérive, maintenir un humain qualifié aux commandes avec autorité de reprise.

Jérôme Vetillard

CTO | VP R&D | Chief Product Officer | AI-Powered Healthcare & Life Sciences Products | Compliance by Design | PhD AgroParisTech | CPO MIT Sloan | Exec MBA IE Business School & Brown University

Twingital-institute / Twingital-ventures : twingital-ventures.com

La meilleure analogie n'est pas le pilote automatique. C'est le médicament. On ne comprend pas toujours son mécanisme d'action au niveau moléculaire. On ne peut pas prédire tous ses effets sur toutes les populations. On le valide cliniquement, on surveille ses effets post-commercialisation, on cartographie ses contre-indications. Personne n'exige qu'il « s'explique » avant d'être prescrit. On exige qu'il soit fiable dans des conditions définies.

De la même manière que calculer l'énergie libre de liaison entre un ligand et son récepteur ne permet pas d'évaluer l'efficacité clinique d'un médicament (les effets d'échelle sont trop nombreux : récepteurs, cellules, tissus, organes, système immunitaire) tracer les états successifs des pondérations d'un réseau de neurones ne permet pas de l'expliquer.

Dans les deux cas, ce sont des phénomènes émergents qui gouvernent. Dans les deux cas, la description du niveau inférieur ne donne pas accès au niveau supérieur. Ce n'est pas une limitation technique. C'est la nature des systèmes complexes.

C'est tout ce qu'on devrait exiger de l'algorithme : non pas l'explicabilité, mais la fiabilité mesurée, auditée, surveillée.

Jérôme Vetillard

CTO | VP R&D | Chief Product Officer | AI-Powered Healthcare & Life Sciences Products | Compliance by Design | PhD AgroParisTech | CPO MIT Sloan | Exec MBA IE Business School & Brown University

Twingital-institute / Twingital-ventures : twingital-ventures.com