# IA DE CONFIANCE :

## enjeux et solutions pour un traitement éthique des données

PAR **VINCENT LUCIANI** (X05), PDG D'ARTEFACT

Les risques liés à l'usage massif de l'intelligence artificielle résident essentiellement dans la reproduction de biais et stéréotypes humains. Pour créer une IA éthique, les développeurs doivent créer *by design* des solutions *data-driven* dans une démarche tant technique que juridique et humaniste.

La science-fiction est pleine d'histoires d'intelligences artificielles qui se révoltent, où les machines agissent selon leur propre système de valeur évidemment maléfique. Souvenez-vous d'HAL 9000 dans *l'Odyssée de l'espace* ou encore de Skynet dans *Terminator* ! De retour dans la vraie vie, les systèmes et algorithmes d'IA ont largement intégré des actes courants de notre vie quotidienne comme le shopping en ligne, la souscription de contrats bancaires ou encore la consultation de contenu sur les réseaux sociaux. Si les cas d'IA rebelle restent un pur fantasme d'auteur, les dérives et dysfonctionnements d'IA sont, eux, devenus courants, souvent avec des conséquences très réelles.

### L'éthique contre les stéréotypes

Nous avons tous vu les images générées par des IA comme Dall-E 2. C'est amusant et innocent de créer une image d'ourson qui écoute de la musique sous l'eau. Mais saviez-vous que ces algorithmes perpétuent des stéréotypes genrés en surreprésentant des hommes dans des images de docteurs, pilotes et PDG, mais en surreprésentant des femmes dans des images d'infirmières, d'hôtesses de l'air, et de secrétaires ?

Vu la masse de décisions du quotidien déléguées à l'IA, nous ne pouvons ignorer les questions éthiques soulevées par son usage. En général, l'IA reste un marché et une technologie peu réglementée encore (la Commission européenne planche activement sur un futur règlement). Il relève donc de la responsabilité individuelle de protéger les utilisateurs.

La problématique de l'IA est de créer *by design* des systèmes de confiance qui répondent aux attentes éthiques de notre société. Il en devient impératif de se préparer, dès maintenant, à concevoir et maintenir des systèmes d'IA de confiance. Ce chantier est vaste, et l'éthique est un sujet qui se construit tout au long du cycle de vie d'un produit.

### Une intelligence artificielle… qui mime les erreurs humaines

L'IA est un ensemble de théories et de techniques permettant à une machine soit de prendre une décision face à une situation prédéfinie, soit de simuler l'intelligence humaine. Cette programmation peut être soit **déterministe**, c'est-à-dire mettant en œuvre une succession de règles humaines à appliquer selon un schéma de décision ; soit **probabiliste**, c'est-à-dire par inférence du bon comportement au regard de l'apprentissage fait sur des situations antérieures réelles, que l'on appelle *machine learning*.

Quel que soit le type de programmation, l'IA est une automatisation de la prise de décision. Les risques éthiques ne résident donc pas dans le fait que l'IA puisse mal automatiser, mais au contraire qu'elle mime parfaitement les décisions humaines, avec son lot d'erreurs et de biais.

Beaucoup de ces dysfonctionnements proviennent du contexte culturel dans lequel a été développée l'IA, lorsqu'elle réplique les systèmes de valeurs historiques d'une société (inégalités de genre ou ethniques, par exemple) ou de ses concepteurs (éducation, sensibilité

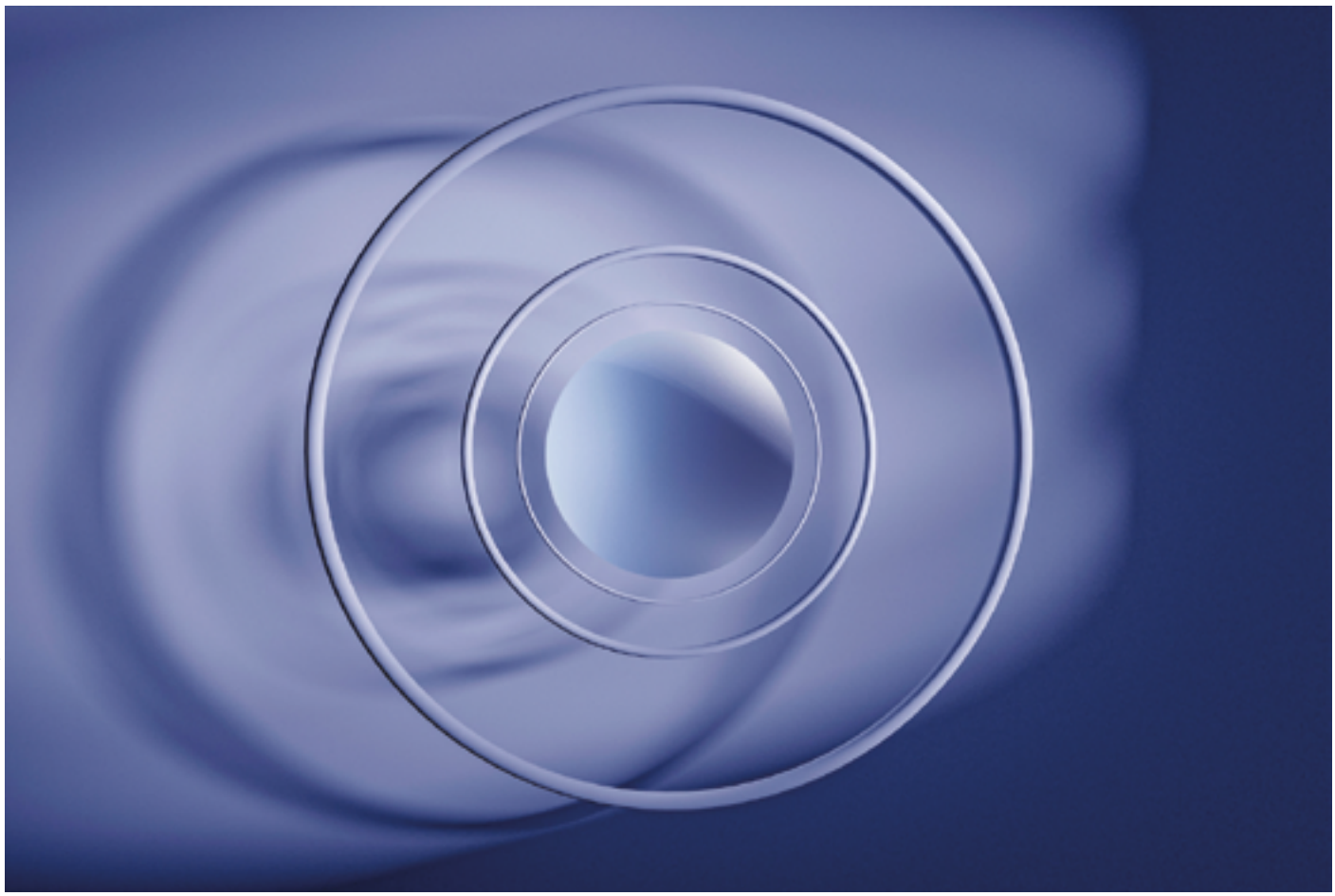


Photo de Milad Fakurian sur Unsplash

politique, religion, etc.). Ces biais sont particulièrement présents dans le cadre des algorithmes de *machine learning* puisqu'ils sont entraînés sur des bases de données historiques, en inférant le futur à partir du passé.

Les chantiers techniques liés à la conception et au cycle de vie d'une IA de confiance répondent tous deux à des questions essentielles telles que : comment détecter avec fiabilité les éventuels biais d'une IA ? Comment interpréter les résultats d'un modèle ? Comment mesurer l'évolution de la performance d'une IA ? Quelles garanties de sécurité pouvons-nous apporter ? Pouvons-nous réagir à une dérive ? Cette liste non exhaustive traduit la diversité des défis à relever pendant la mise en production d'une IA.

## L'IA au secours de l'IA pour corriger ses biais

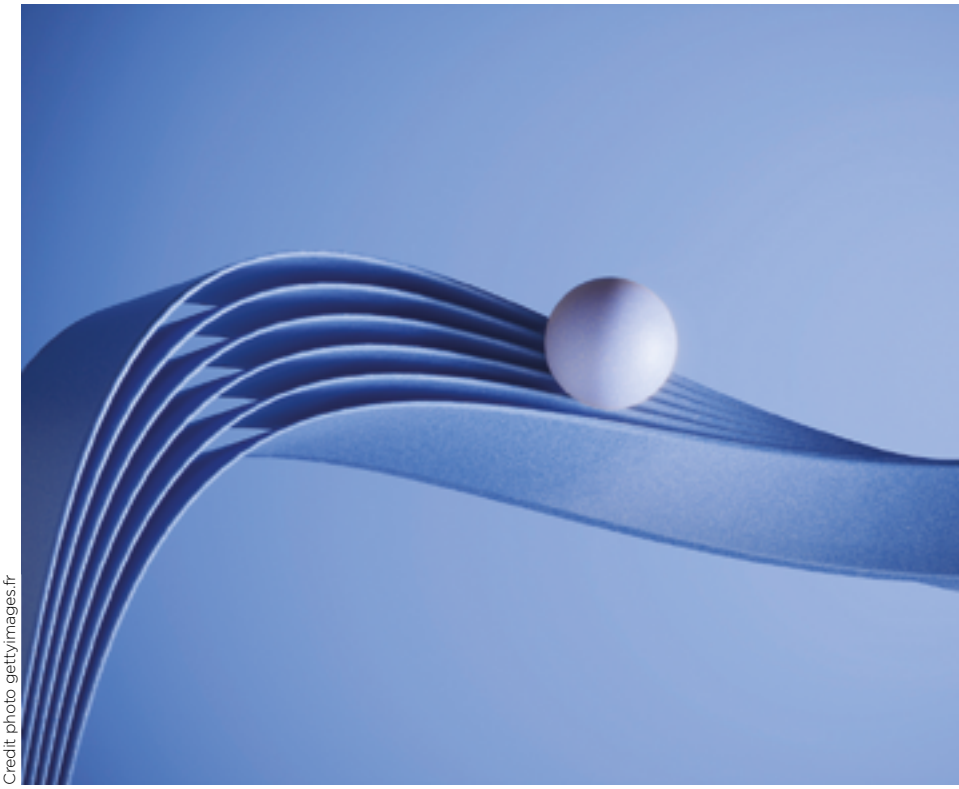
À mesure que les modèles d'IA sont passés de concepts à des produits, nombre de solutions techniques ont été créées pour faciliter l'automation et le déploiement de systèmes IA. Récemment, l'offre de solutions techniques a explosé, créant un écosystème riche et nourri par l'innovation constante de start-up comme de leaders d'industrie. Des exemples notables sont la AI Infrastructure Alliance et la MLOps Community, tous deux dont Artefact fait partie. Trouver le bon outil pour la bonne tâche devient donc plus important, et plus difficile. En parallèle, il y a un écosystème tout aussi vibrant de cabinets de conseil qui s'est développé, pour accompagner leurs clients dans ces choix stratégiques.

> *"Les biais ne viennent pas des* data scientists, *mais des* datasets."

Beaucoup de ces solutions techniques existent en *open source* et sont donc en libre accès à tous. Ces boîtes à outils techniques peuvent détecter et mesurer les biais tout le long de la chaîne de traitement de la donnée : depuis sa collecte jusqu'à son exploitation, en passant par ses transformations et sa modélisation. Elles doivent être activées tout au long du cycle de vie du produit, non seulement en phase de conception et de développement mais aussi en production, afin d'assurer une correction pérenne. Il y a particulièrement trois biais qui peuvent être résolus grâce à des solutions techniques.

## Corriger les biais du passé

Les biais ne viennent pas des *data scientists*, mais des *datasets*. Il faut donc toujours commencer par une exploration et une réflexion profonde des *datasets*. Ces biais peuvent être tant des biais techniques (variable omise, problème de base de données ou de sélection) que des biais de société (économiques, cognitifs, émotionnels).



Credit photo gettyimages.fr

Pour redresser les jeux d'entraînement, il pourrait sembler logique d'effacer toute trace de données sensibles. Pourtant, cela se révèlerait aussi inutile que dangereux : d'une part parce qu'une donnée non sensible peut être insidieusement corrélée à des données sensibles, et d'autre part parce que ce type de raccourcis incite à limiter les contrôles réguliers des données et processus. Pour détecter ces biais, il faut être au courant qu'ils existent et savoir quoi chercher. Il est important de comprendre le contexte des données pour les interpréter correctement. Un programme tout seul ne peut pas identifier de biais, d'où l'importance d'équipes de *data scientists* diverses.

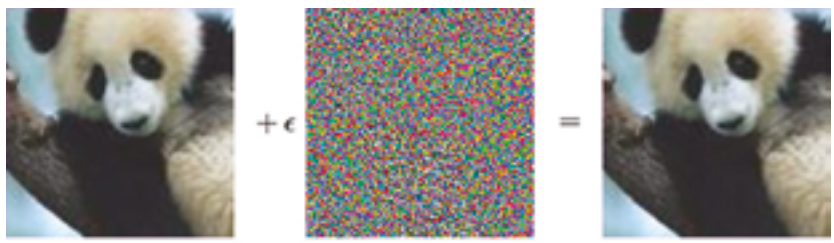
C'est en pensant de façon critique à propos des différentes métriques associées à un modèle d'IA que nous pouvons identifier et rectifier des problèmes. Prenons un modèle qui doit détecter une maladie présente dans 1 % de la population. Si le modèle prédit toujours que la personne est saine, le modèle aura un taux de précision de 99 %. Sans contexte, ce score est excellent. Pourtant, ce modèle est inutile.

L'utilisation de statistiques de bases, univariées et bivariées, du *dataset* peut identifier certains biais. Par exemple, est-ce que chaque groupe d'âge est représenté de manière suffisamment égale ? Une solution pourrait être de changer des données continues en données catégorielles. Des matrices de corrélation peuvent aussi valider des liens entre deux variables liées. Tout ce travail en amont est crucial pour s'assurer que les modèles sont entraînés sur des *datasets* de qualité. En deux mots : *Garbage in, garbage out*.

## Expliquer les résultats d'un modèle

Les modèles d'IA performants tels que les réseaux de neurones artificiels sont très efficaces et utilisés dans de nombreux cas d'usage. Toutefois, ils sont difficilement interprétables. C'est pourquoi ces algorithmes sont aussi appelés boîtes noires (*blackbox models*). En effet, identifier et expliquer la cause des biais plutôt que d'en écarter les conséquences est le grand défi du *machine learning*. Il existe un compromis entre l'explicabilité et la précision. Certains algorithmes comme les arbres de décision sont très explicables, mais moins utiles pour des prédictions complexes.

Le domaine de l'Explainable AI (Ex-AI) essaye de résoudre cette problématique en développant des méthodes et algorithmes qui augmentent l'explicabilité des systèmes. On peut soit explorer la compréhension globale, qui explique comment une IA fonctionne sur la population




globale, ou la compréhension locale, qui explique comment l'algorithme fonctionne sur un exemple en particulier. Cette dernière est demandée par le RGPD car, pour être transparente, une IA doit pouvoir expliquer pourquoi le modèle a pris une certaine décision pour une personne – un objet – une ligne de données en particulier. Il est important de regarder à la fois la compréhension globale et la compréhension locale pour analyser comment le modèle se comporte.

Pour cela, les *data scientists* peuvent regarder les variables d'importance du modèle. Par exemple, il n'est pas normal qu'un modèle de recrutement ait le genre comme variable d'importance. Ils peuvent aussi vérifier les métriques tels que l'*accuracy* (précision : nombre de positifs bien prédits – vrai positif – divisé par l'ensemble des positifs prédits – vrai positif + faux positif), le F1 score (permet

de résumer les valeurs de la précision et du *recall* en une seule métrique) ou le *recall* (nombre de positifs bien prédits – vrai positif – divisé par l'ensemble des positifs – vrai positif + faux négatif), par rapport à une variable (*eg* : la précision pour les femmes *vs* pour les hommes).

## Améliorer la robustesse d'un modèle

Les sources de dysfonctionnement des modèles d'IA peuvent être fortuites, lorsque les données sont corrompues, ou intentionnelles, par des hackers par exemple. Ces deux sources de biais induisent les modèles d'IA en erreur en fournissant des prédictions ou des résultats incorrects.

Certaines solutions *data-driven* permettent d'évaluer, de défendre et de vérifier des modèles et des applications de *machine learning* contre les menaces contradictoires qui pourraient cibler les données (empoisonnement des données), le modèle (fuite de modèle) ou l'infrastructure sous-jacente, tant matérielle que logicielle.

Simplement en superposant une image de « bruit » à une image normale, un classificateur peut être amené à catégoriser à tort un panda comme un gibbon. La différence est imperceptible à l'œil humain, mais cette technique est bien connue comme pouvant tromper des modèles d'IA.

## Développer des systèmes moins humains, mais plus humanistes

Le concept d'IA de confiance ne peut se réduire à sa conception juridique et technique. En effet, le volet relatif à l'humain et à l'organisation est crucial pour mener à bien une démarche éthique liée à l'IA. Les chartes éthiques et les solutions d'amélioration doivent être connues de toutes les parties prenantes, appliquées tout au long du processus de création et suivies dans le temps. Cela nécessite une transformation de la culture de l'organisation pour y intégrer en profondeur les thématiques et approches éthiques, pour garantir la pérennité des solutions et contribuer à l'IA éthique *by design*.

L'IA elle-même n'est pas éthique ou non éthique. Ce sont uniquement la manière dont on a entraîné le système et la manière dont on s'en sert qui sont éthiques ou pas. Le véritable risque éthique lié à l'usage massif de l'IA n'est donc pas que les algorithmes se révoltent. Au contraire, le risque intervient précisément lorsque l'IA se comporte exactement comme nous l'avons demandé, mimant nos biais, répétant nos erreurs, amplifiant nos incertitudes et nos imprécisions.

> *"Le risque intervient précisément lorsque l'IA se comporte exactement comme nous l'avons demandé."*

## Soyons proactifs sur le sujet de l'IA éthique

C'est un sujet qui nous tient à cœur, à Artefact, et c'est pour cela que nous y consacrons des équipes, et que nous avons développé un accompagnement spécifique sur ces enjeux d'élaboration de gouvernance, d'implémentation de solutions techniques et de conseil de stratégie IA. Nous travaillons en étroite collaboration avec le monde académique, étant partenaire de la chaire Good in Tech, cofondée par l'Institut Mines-Télécom et Sciences Po, pour réduire l'écart entre recherche et pratique. Artefact a aussi reçu le label Responsible and Trusted AI, décerné par l'association indépendante Labelia Labs, qui garantit un haut niveau de maturité sur les sujets d'IA responsable et de confiance.

La question de l'éthique nous concerne tous et il convient donc d'être proactif sur le sujet. Adopter un comportement éthique ne doit pas seulement être une réaction. L'intelligence artificielle traverse une période importante, où ses procédés et son ethos sont en pleine phase de définition. Dans un domaine pas encore régulé, il revient à celles et ceux qui y prennent part de prendre les devants. C'est notre responsabilité à tous, en tant que développeur, consultant ou manager, de répondre aux attentes des personnes concernées – clients et utilisateurs des services, mais surtout la société en général – et d'utiliser les pouvoirs de l'IA pour créer un monde meilleur.