IMPACT AI

LES BRIEFS DE L'IA RESPONSABLE

3

IA générative et transparence

NOVEMBRE 2024

Introduction

« Toute technologie suffisamment avancée est indiscernable de la magie »

Arthur C. Clarke

L'IA générative est sans aucun doute une technologie innovante et transformatrice, mais elle nécessite une compréhension approfondie de la part des utilisateurs (et éventuellement d'autres parties prenantes) pour qu'elle apporte toute sa valeur... tout en restant maîtrisable.

Ce nouveau « Brief de l'IA responsable » d'Impact AI rend compte de nos travaux et propose des pratiques et des initiatives visant à s'assurer que l'utilisation de l'IA générative (IA Gen) offre une transparence maximale.

Il s'appuie sur les retours d'expérience des membres d'Impact AI et sur plusieurs études, rapports et autres initiatives récents, afin de fournir un aperçu des stratégies visant à atténuer la dépendance excessive à l'IA et à maximiser l'intelligibilité.

1/ Le contexte

Dans une situation de fort développement des utilisations liées à l'IA Gen, la transparence mais aussi l'explicabilité d'une réponse ou d'un résultat ont de multiples facettes. On peut citer la transparence sur les processus d'apprentissage, sur les données utilisées pour l'entraînement et le test.
La transparence consiste également à déterminer comment les parties prenantes (directes ou indirectes) pourraient mal comprendre, mal utiliser ou estimer de manière incorrecte les capacités, les limites et les résultats d'un système d'IA Gen. Le but est de leur permettre une compréhension adéquate du fonctionnement du système. Cependant, ce qui rend un tel système « intelligible » est difficile à cerner. Concept fondamentalement humain, l'intelligibilité manque d'une définition universelle. Le niveau d'intelligibilité est évidemment fonction des compétences des parties prenantes, de leurs caractéristiques démographiques, professionnelles, sociales et culturelles, de leur expertise du domaine, de la fréquence de leurs interactions avec le modèle sous-jacent, de leur familiarité avec le système, etc. L'intelligibilité est également étroitement liée à la capacité des utilisateurs de former des modèles mentaux, qui reflètent la réalité complexe du système considéré.
Bien que de nombreuses approches techniques de l'intelligibilité aient été exposées dans la littérature sur l'apprentissage machine (ML), l'intelligibilité de l'IA Gen est encore un domaine relativement inexploré.

Confiance vs dépendance excessive

Étroitement liée à l'intelligibilité, la dépendance excessive à l'IA Gen se manifeste lorsque les utilisateurs commencent à accepter des sorties incorrectes et estiment qu'elles sont basées sur des faits alors qu'en réalité elles ne sont pas exactes. Il existe plusieurs mécanismes sous-jacents associés (biais d'automatisation, biais de confirmation, effets d'ordre et surestimation des explications) et leurs conséquences peuvent varier considérablement, allant des moins graves, s'il s'agit de l'écriture d'une annonce publicitaire, à très graves s'il s'agit d'un diagnostic médical.
Créer le bon niveau de confiance dans les systèmes d'IA Gen est un processus complexe. Ces systèmes présentent en effet un certain nombre de risques, même s'ils rivalisent souvent, et parfois surpassent, les performances humaines dans de nombreuses tâches.
Un niveau de confiance inapproprié dans l'IA Gen – sous-estimé ou excessive – peut entraîner des conséquences néfastes, en particulier dans le cas où les performances du couple humain/IA sont moins bonnes que celles qui auraient été obtenues si l'humain et l'IA avaient travaillé chacun de leur côté.
Un objectif important de la conception d'un système d'IA Gen est de permettre aux utilisateurs d'accorder le bon niveau de confiance à l'IA et cela se produit lorsque les utilisateurs acceptent les résultats d'IA corrects et rejettent les résultats incorrects.
La réglementation européenne sur l'IA (EU AI Act) impose d'ailleurs des dispositifs de contrôle humain, ce qui fait des utilisateurs la dernière ligne de défense contre les défaillances des systèmes d'IA.

Transparence entre les différents acteurs de la chaîne de valeur

La chaîne de valeur d'un système d'IA, de son concepteur aux utilisateurs, peut parfois être complexe et faire intervenir des acteurs multiples. La transparence dans ces relations permettra, par exemple, de préciser clairement les responsabilités de chacun, leurs contributions telles que la documentation du modèle d'IA Gen et du système.
Et en bout de chaîne, la transparence vis-à-vis des utilisateurs de solutions permet de créer le bon niveau de confiance en l'IA, en identifiant clairement une interaction avec un humain ou une IA, un contenu généré par une IA Gen, ou les limites de chaque IA.

2/ Les retours d'expérience des membres d'Impact AI :

Microsoft France

Microsoft est engagé sur la voie d'une IA responsable, c'est à dire sûre, sécurisée et transparente, depuis plus de huit années. L'entreprise a partagé récemment ses apprentissages sur l'évolution de ses pratiques d'IA responsable pour relever les nouveaux défis soulevés par l'IA générative, dans un premier rapport annuel de transparence de l'IA responsable, qui sera publié chaque année.

Celui-ci révèle la priorité donnée à :

→ **Respecter ses principes d'IA pour le développement et l'utilisation responsables de l'IA (générative) au sein de Microsoft**. Avec l'IA générative, le principe de Transparence a amené à investir dans de nouvelles approches en vue d'une expérience utilisateur intelligible, d'explications pertinentes et d'une vérification facilitée des résultats, etc.

→ **Travailler avec les leaders de l'industrie et les gouvernements** afin d'élaborer de nouvelles normes pour les modèles de fondation à haute capacité. Cela se traduit, notamment, par le soutien des travaux du Frontier Model Forum avec le partage de connaissances, le développement de meilleures pratiques et en faisant progresser la recherche dans le domaine de l'IA.

→ **Donner à nos clients et partenaires les moyens de développer et d'utiliser l'IA de manière responsable**, avec notamment la publication de 33 notes de transparence depuis 2019, pour permettre une utilisation et une intégration responsables de ses services de plateforme d'IA, ou encore le lancement de 30 outils d'IA responsables avec plus de 100 fonctionnalités. Ces outils incluent la boîte à outils HAX (Human-AI eXperience) pour une IA centrée sur l'humain et le SDK d'évaluation Azure AI vis-à-vis des contenus générés.

« Notre engagement en faveur d'une IA responsable donne la priorité à maintenir les humains non seulement dans la boucle, mais aussi au centre des systèmes d'IA. »

Natasha CRAMPTON
Vice President
et Chief Responsible AI Officer,
Microsoft Corporation

AI-vidence

Ai-vidence est une start-up dédiée à l'explicabilité des IA pour leur adoption ou leur substitution par des modèles plus fiables. De nos travaux pour le collectif Confiance.IA, où l'utilisation autonome des systèmes nécessite de contrôler leur robustesse, et pour le régulateur financier, où un contrôle humain est indispensable, nous retenons quelques concepts clef pour répondre opérationnellement aux enjeux de transparence. Parmi ceux-là :

→ Adapter l'interface à chaque interlocuteur : sa culture, et ses enjeux métier.

→ Réduire la complexité au niveau adapté aux exigences de chaque interlocuteur.

→ Veiller à ce que les parties prenantes soient impliquées suffisamment en amont dans toutes les phases projet d'élaboration.

D'un point de vue méthodologique, dans ces deux contextes différents, requérant robustesse ou neutralité technologique, il est également bienvenu d'opter pour une démarche multidisciplinaire, itérative, afin d'assurer que le rendu, interface et contenu, soit en ligne avec les attentes de l'expert métier. Car c'est lui le point focal, qui devra : renseigner les dossiers pour les régulateurs, estimer les risques effectifs pour l'entreprise (comme pour le MRM en banques), et expliquer au besoin à un client final la logique d'un traitement. Nous privilégions une approche multi-échelles, afin de proposer le bon niveau entre synthèse et analyse (pour un élément industriel, apporter une réponse par exemple au niveau du site industriel, mais également du sous-système technico-fonctionnel, etc.) et pour contextualiser la justification apportée (un même élément ou dossier client pouvant relever d'une logique très différente selon l'endroit ou le segment de clientèle pour lequel l'IA est utilisée).
D'un point de vue technologique, le défi repose également dans la complexité et les intrications de ces systèmes. Il est particulièrement aigu lorsque les risques sont extrêmes, comme dans les systèmes d'arme, pour lesquels réactivité, simplicité des interfaces et fiabilité des informations sont trois composantes difficiles à concilier, tout particulièrement lorsqu'une partie des informations existe sous la forme de schémas logiques, ou de tableaux de formes diverses.
Pour cela, nous utilisons et développons des outils spécifiques à ces besoins :

→ d'exploration et substitution assistées des modèles d'IA par des modèles plus simples et frugaux.

→ de comparaison des données d'entraînement et des données réelles, de leur dérive dans le temps .

→ de découpe assistée en divers changements de contextes.

→ des approches causales, les plus explicables et correctibles.

Tout ceci permet la critique constructive des logiques de fonctionnement de l'IA, leur adoption et leur mise sous contrôle afin d'anticiper les besoins futurs de remédiation, dès les premiers contrôles en août 2026.

Carrefour

Les LLMs sont entrés extrêmement rapidement dans le quotidien personnel, mais aussi professionnel de beaucoup de nos collaborateurs, sans formation préalable. Un important travail d'information sur leur fonctionnement et de sensibilisation sur leurs limites (notamment les biais) et leurs risques, doit être conduit. Comprendre les principes généraux de fonctionnement, notamment quelques concepts rudimentaires d'apprentissage statistique, présente un double bénéfice pour les utilisateurs :

→ **En désacralisant cet outil « magique », cela permet de l'utiliser avec un regard critique, plus prudent.**

→ **Cela permet également aux utilisateurs de mieux interpréter les résultats renvoyés par les LLMs, et donc d'en avoir progressivement un usage plus performant.**

Les LLMs sont de moins en moins utilisés seuls, mais au sein de logiciels intelligents, aussi appelés agents, qui vont pouvoir exécuter des séquences de prompts, mais aussi d'autres tâches. Ces agents sont une opportunité pour pallier les défauts et les risques d'un LLM « brut ». En outre, il est désormais possible de demander à l'agent de vérifier une information auprès d'outils extérieurs, comme une base de données, de le contraindre à citer ses sources ou bien de suivre les étapes d'un raisonnement préétabli.

Par leur comportement stochastique, on peut s'interroger sur la possibilité d'intégrer des LLMs dans des systèmes industriels. Mais ce caractère stochastique n'est pas une nouveauté propre aux LLMs. De nombreux systèmes industriels antérieurs aux LLMs reposaient déjà sur des algorithmes d'apprentissage automatique stochastiques (que ce soit à l'apprentissage ou à l'inférence). Pour opérer des algorithmes d'apprentissage automatique de façon sécurisée en production, il est important d'adopter une approche MLOps : versionnement du code et des modèles, tests unitaires et fonctionnels, déploiement automatisés, évaluation de la performance, surveillance de la dérive des données…

Dans le cas des modèles de fondation comme les LLMs, l'entraînement des modèles n'est souvent plus effectué par les équipes qui vont les exploiter. Mais cette approche MLOps demeure nécessaire pour l'opérer. Elle permet par exemple de tester efficacement de nouvelles versions de LLMs, proposées régulièrement par les fournisseurs.
Le développement d'agents est certes devenu abordable pour un plus large public, du fait qu'une grande partie du développement se fait désormais en langage naturel. Mais si l'on a pour objectif d'industrialiser des solutions, il reste indispensable d'adopter les bonnes pratiques de développement logiciel et de data science, notamment la mise en place de protocoles rigoureux de mesure de la performance.
Ce sont là des conditions nécessaires mais pas suffisantes à l'explicabilité de ces systèmes.

3/ Les meilleures pratiques identifiées

À l'instar de l'IA traditionnelle, il est important d'adopter une stratégie centrée sur l'humain lors de la conception de techniques de transparence, d'intelligibilité ou lors de la vérification que ces techniques atteignent les objectifs visés.
On peut également soutenir que les notions de transparence et d'intelligibilité devraient être étendues, au-delà des modèles d'IA Gen, à d'autres composants ou couches du système.
L'accent est mis ici sur la façon dont les utilisateurs finaux interagissent avec le système IA Gen et sur la façon de créer une expérience adéquate, qui aide les utilisateurs à :

1. Comprendre et utiliser efficacement la technologie IA Gen tout en évitant les pièges communs

2. Renforcer la responsabilité des utilisateurs

3. Mettre en évidence les limites de l'IA Gen pour atténuer la dépendance excessive

4. Enfin, et surtout, s'assurer que les utilisateurs sont conscients qu'ils interagissent avec une IA

Quelques suggestions pour travailler en confiance avec l'IA Gen :

→ **Prenez en considération, avant toute chose, le niveau de transparence souhaité**, qui peut guider le choix de la technologie la plus appropriée. Une solution exigeant une pleine explicabilité de toutes ses décisions ne semble pas actuellement éligible à l'utilisation de l'IA Gen, et devrait conduire à une autre technologie.

→ **Informez les utilisateurs qu'ils interagissent avec un système IA Gen** (par opposition à un autre humain). Le cas échéant, informer les consommateurs de contenus en aval, du fait que ces contenus ont été partiellement ou entièrement générés par un modèle d'IA Gen ; ces informations peuvent être exigées par la législation ou les meilleures pratiques applicables, peuvent réduire la dépendance inappropriée aux résultats générés par l'IA et aider les personnes à exercer leur propre jugement sur la façon d'interpréter et d'agir sur ce contenu.

→ **Structurez les entrées et/ou les sorties du système**. Utilisez des techniques de « prompt engineering » au sein du système d'IA Gen pour structurer les entrées dans le système afin d'éviter les questions/réponses ouvertes.
Les sorties peuvent également être structurées et limitées dans certains formats ou modèles avec le message système et/ou le « prompt ».
Par exemple, si le système d'IA Gen génère un dialogue pour un caractère fictif en réponse à des requêtes, limitez les entrées de sorte que les utilisateurs puissent uniquement interroger un ensemble de concepts prédéterminés.

→ **Citez les références et les sources d'information**. Si le système d'IA Gen génère du contenu en se basant sur les références envoyées au modèle, le fait de citer clairement les sources d'information aide les personnes à comprendre d'où provient le contenu généré par l'IA.

→ **Mettez en évidence les inexactitudes potentielles dans les résultats générés par l'IA,** à la fois lorsque les utilisateurs commencent à utiliser le système IA Gen

et à des moments appropriés pendant l'utilisation continue. Dès la première exécution, avertissez les utilisateurs que les sorties générées par l'IA peuvent contenir des inexactitudes et qu'ils doivent vérifier soigneusement les informations. Tout au long de l'expérience, incluez des rappels et points de contrôles pour vérifier les résultats générés par l'IA pour des inexactitudes potentielles, à la fois globales et par rapport à des types spécifiques de contenu que le système pourrait générer incorrectement.

→ **Revoyez et modifiez les résultats générés : concevez l'expérience utilisateur (UX),** pour encourager les personnes qui utilisent le système d'IA Gen, à revoir et modifier les résultats générés par l'IA avant de les accepter.

→ **Empêchez l'anthropomorphisation du système.** Les modèles d'IA Gen peuvent produire du contenu contenant des opinions, des déclarations émotives ou d'autres formulations qui pourraient sous-entendre qu'ils sont semblables à des êtres humains, être confondus avec une identité humaine ou induire les personnes en erreur en leur faisant croire qu'un système d'IA Gen a certaines capacités, alors que ce n'est pas le cas. Mettez en œuvre des mécanismes qui réduisent le risque de tels résultats ou intégrez des informations à fournir, pour éviter toute interprétation erronée des résultats.

→ **Responsabilisez l'utilisateur.** Rappelez aux utilisateurs qu'ils sont responsables du contenu final lorsqu'ils examinent du contenu généré par l'IA.

En outre, le système d'IA Gen devrait être positionné de manière adéquate, le cas échéant, dans le contexte :

→ **Informez et accompagnez les utilisateurs.** Produisez et fournissez du matériel pédagogique pour le système d'IA Gen, y compris des explications sur ses capacités et ses limitations, les utilisations prévues, la plage de confiance éventuelle, etc. Par exemple, cela pourrait prendre la forme d'une page « en savoir plus », accessible via l'application.

→ **Publiez des directives utilisateur et des meilleures pratiques.** Aidez les utilisateurs et les autres parties prenantes à utiliser le système d'IA générative de manière appropriée en publiant les meilleures pratiques, par exemple en matière d'élaboration rapide, en passant en revue les générations avant de les accepter, etc. Ces lignes directrices peuvent aider les personnes à comprendre comment fonctionne le système d'IA Gen. Dans la mesure du possible, incorporez les directives et les meilleures pratiques directement dans l'UX.

→ **Visez le niveau de transparence adéquat.** Il est important d'offrir le bon niveau de transparence aux personnes qui utilisent le système d'IA Gen, afin qu'elles puissent prendre des décisions éclairées concernant l'utilisation du système ainsi que la pertinence et l'applicabilité des sorties générées par l'IA. Cette bonne pratique reste, comme nous l'avons vu, un défi nécessitant de poursuivre les travaux.

4/ Les défis et questions ouvertes

Bien que prometteuse, l'IA Gen en est encore à ses balbutiements s'agissant de son appropriation par les utilisateurs. Ce sont des techniques complexes. Leur efficacité dépend du contexte d'utilisation, des tâches à accomplir. Il est donc nécessaire de se fixer des attentes réalistes et lourdement testées.

De même, il faut admettre que l'ensemble des suggestions émises dans ce Brief de l'IA responsable d'Impact AI ne sont pas appropriées pour chaque scénario d'utilisation de l'IA générative et qu'inversement, elles peuvent être insuffisantes dans certains scénarios spécifiques.

5 / Pour aller au-delà

• **Overreliance on AI : Literature Review**, un rapport qui explique en quoi consiste la dépendance excessive à l'IA, comment elle se produit et comment nous pouvons l'atténuer. Il montre comment et pourquoi une dépendance excessive à l'IA fait qu'il est difficile pour les utilisateurs d'exploiter de manière significative les forces des systèmes d'IA et de surveiller leurs faiblesses.

Basé sur une revue de la littérature de près de 60 articles de différents domaines de recherche, ce rapport fournit un aperçu détaillé de la façon dont la dépendance excessive à l'IA se produit, comment mesurer la dépendance excessive, quelles sont ses conséquences et comment nous pouvons minimiser ses effets négatifs.

• **Confiance appropriée à l'IA générative : synthèse de la recherche**, un rapport qui fournit un aperçu des facteurs qui affectent la dépendance excessive à l'IA Gen, l'efficacité des différentes stratégies d'atténuation de la dépendance excessive à l'IA Gen, et les stratégies de conception potentielles pour faciliter une confiance appropriée à l'IA Gen. Ce rapport est basé sur l'examen d'une cinquantaine d'articles provenant de plusieurs domaines de recherche.

• Human-AI eXperience (HAX) Toolkit, un ensemble d'outils pratiques pour créer des expériences humains-IA. Chaque outil s'appuie sur les besoins observés, validés par des recherches rigoureuses, puis est testé avec des équipes de praticiens. La boîte à outils HAX comprend actuellement 18 directives dans la bibliothèque de conception HAX pour planifier une application d'IA (Gen), le manuel HAX Workbook pour prioriser les directives à appliquer et comment les mettre en œuvre, et à cette fin, des exemples et des modèles de conception pour mettre en œuvre chaque directive (pertinente), etc.

• Interprétabilité - **Recherche sur la compréhension du fonctionnement interne des modèles d'IA (identification des concepts)** au sein de Claude 3.5 Sonnet.

www.impact-ai.fr
contact@impact-ai.fr