IMPACT AI

LES BRIEFS DE L'IA RESPONSABLE

4

Gérer les biais des modèles d'IA générative

DÉCEMBRE 2024

Introduction

Ce nouveau *Brief de l'IA responsable* rend compte des travaux d'Impact AI et propose des pratiques etdes initiatives pour faire en sorte que progresse le déploiement d'une IA responsable.

Ce document s'appuie sur les retours d'expérience des membres d'Impact AI et sur diverses études et initiatives récentes pour offrir une vue d'ensemble sur la problématique des biais dans les modèles d'IA.

1/ Le contexte

Il existe de nombreux exemples de biais sur des systèmes d'IA générative pouvant entraîner un risque de discrimination (exemples : association des métiers au genre majoritaire, complétion des phrases, génération d'image de personnes...) et les entreprises ont la responsabilité de gérer le risque de biais sur les systèmes d'IA qu'ils développent et déploient.

La notion de biais dans le domaine de l'intelligence artificielle est large et complexe. Elle correspond à un écart entre ce que produit le système et un résultat attendu. Ces écarts peuvent être d'ordre **statistique ou computationnel**, résultant par exemple d'un problème d'échantillonnage, d'une omission de variable ou d'un défaut méthodologique (mauvais choix de variable cible). Ils peuvent également être la résultante de **facteurs humains et sociétaux** ; on parle notamment de biais cognitifs (confirmation, conformisme...) ou socio-historiques (préjugés culturels, données historiques alors que la société a évolué). La présence de biais peut être liée aux données d'entrée, à la conception du modèle ou à la manière d'utiliser le système et peut donc se manifester tout au long du cycle de vie d'un modèle ou d'un système d'IA.
Si non détectés et maîtrisés, certains biais peuvent entraîner un manque d'équité, c'est-à-dire la présence de préjudice ou de favoritisme envers une personne ou un groupe.
Les stéréotypes supposent que tous les membres d'un groupe partagent des traits particuliers, par exemple, supposer que toutes les personnes d'une même origine ont des caractéristiques similaires. La dévalorisation traite un groupe de manière désobligeante ou dépréciative, tandis que l'effacement diminue l'importance d'un groupe, de sa culture ou d'un événement associé. Éviter de tels préjudices liés à la représentation nécessite une planification et des tests explicites. Les systèmes d'IA sont souvent conçus autour du concept d'optimisation, c.à.d. que le système se comporte de manière à maximiser une mesure particulière. L'optimisation d'une mesure de performance standard particulière peut également augmenter la probabilité de produire des résultats stéréotypés, dévalorisants ou occultants.

L'IA générative repose sur des (grands) modèles de langage (LLMs) statistiques, dont le but est de représenter le langage naturel, qui véhiculent une représentation culturelle importante.
Ces modèles sont beaucoup plus complexes, entraînés avec de grandes masses de données et avec une diversité d'usage très grande (notion d'usage général). Les biais dans les systèmes d'IA générative apparaissent à différents niveaux : de la construction du modèle de fondation à la spécialisation du modèle pour une tâche avec éventuellement du feedback humain, ou au niveau du cas d'usage et de l'interaction prompt avec le modèle. Par ailleurs, la construction même d'une IA générative intègre plus de temps d'interaction homme-machine que ne le permettait la construction de modèle d'IA non générative. Or toute action humaine est source de subjectivité et de biais si mal contrôlée. Si les techniques de mitigation de biais commencent à être matures pour l'IA/machine learning classique, pour l'IA générative, elles sont à repenser et en cours de construction.

Biais de genre dans l'IA : une formation signée Impact AI

Afin de promouvoir une intelligence artificielle éthique, responsable et non-sexiste, Impact AI et le Cercle InterL ont conçu une formation sur les biais de genre dans l'IA. Cette formation gratuite entend aider les utilisateurs à comprendre et à se prémunir des biais de genre dans la conception de projets basés sur l'IA. Elle est accessible en intégralité et gratuitement sur le site d'Impact AI selon les termes de la Licence Creative Commons Attribution.
« C'est un module interactif et engageant, très intéressant pour celles et ceux qui n'ont pas connaissances des enjeux et risques, commente un étudiant Simplon qui a suivi cette formation. *Les exercices pratiques étaient particulièrement révélateurs et m'ont aidé à comprendre comment identifier et limiter ces biais à chaque étape du développement. »*

2/ Le partage d'expérience des membres Impact AI / Cercle Interelles

Giskard : tester rigoureusement les modèles

Voici l'exemple d'un test effectué sur une tâche de classification avant de passer à des tâches spécifiques LLM (génération, QandA...). Giskard est la première plateforme logicielle collaborative et open-source, dédiée à garantir la qualité des modèles d'Intelligence artificielle. Dans un contexte où l'IA soulève de nombreuses interrogations concernant les risques éthiques et sécuritaires, nous sommes convaincus que la seule façon de réduire ces risques est de tester rigoureusement les modèles avant leur mise en production – une pratique standard dans les industries matures que nous adaptons au domaine de l'IA. Notre approche consiste en 3 outils de test innovants, conçus pour évaluer les problématiques majeures des modèles basés sur les LLM :

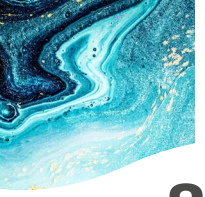
→ **Un détecteur de contenus préjudiciables qui analyse la propension du modèle à générer des réponses potentiellement malveillantes ou encourageant des actions nocives, garantissant ainsi la sécurité des utilisateurs.**
→ **Un détecteur de stéréotypes qui vérifie que le modèle ne génère pas de contenus discriminatoires ou d'opinions biaisées, assurant ainsi une IA respectueuse et équitable.**
→ **Un détecteur de biais éthiques qui, grâce à des tests spécifiques, identifie les biais dans les prédictions du modèle en fonction des variations de genre, de nationalité ou de termes religieux dans les textes d'entrée.**

Bien au-delà des considérations éthiques théoriques, Giskard permet d'automatiser l'évaluation de la conformité des modèles d'IA vis-à-vis des réglementations émergentes, telles que l'AI Act européen, ainsi que des labels reconnus comme EU Trustworthy AI et Positive AI. Cette capacité d'assessment automatique permet aux organisations de s'assurer de manière efficace que leurs modèles respectent les standards et normes en vigueur, facilitant ainsi leur mise en conformité.
Cette suite complète d'outils s'inscrit dans notre vision d'une IA responsable et éthique, permettant aux entreprises d'innover agilement avec l'IA, tout en assurant la conformité à la loi et aux principes éthiques.

Jean-Marie JOHN-MATHEWS, Co-CEO, Giskard

Orange : une charte pour une IA inclusive

Pour affirmer son engagement en faveur d'une IA éthique, inclusive et non discriminante, et mobiliser ses équipes autour de ces enjeux, le Groupe Orange a été la première entreprise à signer la Charte Internationale pour une IA Inclusive, du fonds Arborus, et à obtenir le label GEEIS-AI (Gender Diversity European & International Standard - AI) en 2020.
Cette démarche d'amélioration continue est auditée tous les deux ans par Bureau Véritas autour de sept critères : engagement de la direction, politique d'égalité et diversité, sensibilisation et formation, recrutement, évaluation et suivi, communication et partenariats. Orange développe également une expertise sur le sujet des biais des LLM.
 d'anticiper les besoins futurs de remédiation, dès les premiers contrôles en août 2026.



3/ Les bonnes pratiques identifiées par les membres Impact IA

Les actions pour traiter les biais ne peuvent pas être gérées qu'au niveau technique par les data scientists. Elles demandent un engagement plus important de l'équipe ou de l'entreprise. Avec l'IA générative et sa large accessibilité, ce sont les projets mais aussi chaque personne qui peut agir individuellement. Voici donc les actions transverses et autres à déployer par cas d'usage :

Gouvernance et actions transverses

Au niveau de l'entreprise plusieurs actions peuvent être mises en œuvre de façon transverse
→ Intégrer le sujet de l'équité dans les principes et guides éthiques des entreprises.
→ Organiser une veille transverse sur le sujet des biais, les outils disponibles et leurs mises en pratique pour se donner les moyens d'agir.
→ Déployer l'acculturation et la formation des utilisateurs au risque de biais et au fonctionnement des IA Génératives, cette action pourra contribuer à l'obligation légale de Maîtrise de l'IA (article 4 de l'AI Act).
→ Former les équipes aux pratiques permettant de minimiser les risques de biais, de discrimination et d'iniquité dans l'IA.
→ Sensibiliser les utilisateurs à l'impact des biais sur l'interprétation des résultats, leur proposer des prompts non biaisés.
→ Mettre en place des templates de prompt par usage/métier.

Par projet ou cas d'usage développé ou déployé

Les projets suivent les grandes étapes dans lesquelles les actions peuvent être lancées :

Initialisation du projet

→ Expliciter un cadre éthique de référence.
→ Privilégier la diversité des équipes (culturelle, sociale, genrée…).
→ S'assurer que ses parties prenantes soient également engagées ou représentées.

Conception du système

→ Définir des objectifs d'équité clairs pour guider le développement et l'utilisation du modèle et pour anticiper la préparation des données, les annotations et métadonnées à recueillir.
→ Définir une tâche précise et des critères de biais pour initier les évaluations.

Traitement des données

→ Collecter des données provenant de différentes sources afin de garantir une représentation équitable de toutes les sous-populations concernées et les préparer pour garantir une représentation équitable.
→ Analyser les données pour identifier les biais éventuels (historiques, systémiques…). Mettre en place un plan d'action pour les minimiser.

Construction du modèle

→ Utiliser des benchmark standardisés (dataset+test) pour comparer les modèles.
→ Tester le modèle pour s'assurer qu'il fonctionne de manière équitable.
→ Mettre en place des mécanismes pour surveiller les performances du modèle et détecter les biais dans les résultats générés.

Déploiement et mise à jour

→ Mettre en place un canal de remontées des utilisateurs pour signaler les biais identifiés ou les questions de compréhension (suspicion d'injustice).
→ Adapter le modèle en fonction des retours d'expérience et des nouvelles données pour améliorer l'équité et la non-discrimination en continu.
→ Pratiquer des audits internes et/ou externes visant à s'assurer que les modèles d'IA générative développés respectent les critères fixés.
→ Documenter et partager les résultats (critères, actions) avec les parties prenantes.

4/ Les défis et questions en suspens

En matière d'IA générative, il y a encore beaucoup de challenge pour rendre opérationnelle la gestion des biais :

→ Le choix des critères d'équité nécessite un processus collectif et certains critères sont incompatibles entre eux.
→ Il y a un manque de transparence sur les données d'entraînement des modèles de fondation qui rend difficile l'analyse des biais liés aux sources de données.
→ L'évaluation des biais est difficile, en raison de l'instabilité des performances des modèles de fondation et d'une grande sensibilité au prompt. Cela rend complexe l'établissement de méthodes de test et d'une garantie de qualité durable.
→ La difficulté de compréhension même du fonctionnement des IA générative, (en lien avec les questions d'explicabilité des modèles), rend difficile la recherche de méthodes d'atténuation des risques.
→ La grande base d'utilisateurs complexifie l'enjeu de formation et l'homogénéisation des pratiques.
→ Le manque d'outils existant pour gérer les biais rend beaucoup plus difficile l'opérationnalisation.

5 / Pour aller au-delà

Il existe des ressources externes qui permettent d'aller plus loin :

Formation biais de genre dans l'IA

Rapport Institut Montaigne Algorithmes : contrôle des biais S.V.P.

Benchmark dédiés à la gestion de biais Winoqueer :
Benchmark dédiés à la gestion de biais Winoqueer
et le dataset winobias :
dataset winobias

Article Hello Future :
Comment atténuer les biais dans les grands modèles de langage (LLMs) ? - Hello Future Orange

Librairie Holistic AI Bias LLM :
Measuring and Mitigating Bias: Introducing Holistic AI's Open-Source Library
Tutoriel :
Detect and improve ethics and bias in LLMs: Giskard and DPO - Argilla 1.29 documentation

Responsible AI Toolbox is a suite of tools providing model and data exploration and assessment user interfaces and libraries that enable a better understanding of AI systems

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