L’Intelligence Artificielle est un homme blanc Américain, trentenaire..

..et probablement protestant

Je manipule l’Intelligence Artificielle au quotidien, en particulier ChatGPT, Claude pour la production de textes et avec lesquels je construis et entraine mes propres modèles. Et Midjourney qui me sert à travailler sur mes propres images et illustrations.

Avec une formation en linguistique et sémiologie (sous la direction de Dominique Maingueneau, cela a son importance ici dans l’analyse du dicours car l’image est aussi un discours) et en informatique, l’IA est selon moi l’illustration parfaite du lien qui existe entre ces deux disciplines.

L’idée de cette expérience m’est venue en parcourant différentes images générées par l’IA. J’ai remarqué que peu d’images montraient un travail approfondi sur les prompts, et que celles qui étaient générées paraissaient souvent sans vie. Certaines images se détachent clairement du lot de banalités : leurs auteurs, comme je le pratique, introduisent dans l’IA leurs éléments graphiques (personnage, texture, compositions), connaissent les règles de composition de l’image et élaborent des prompts complexes pour parvenir au résultat qui correspond aux attentes. MidJourney offre d’immenses possibilités de créations dépassant le simple cadre d’image générées aléatoirement par IA.

Par ce biais, on retrouve un peu cette époque de la Renaissance où les apprentis recevaient les consignes et exécutaient l’oeuvre de leur Maître.

Mais c’est une image banale, toutefois accompagnée d’un prompt complet et plutôt élaboré, que j’ai choisie pour cette expérience, car en regard du prompt qui l’avait produite, elle accumulait selon moi de nombreux biais et en premier lieu un biais racialiste.

L’IA, telle qu’elle est souvent déployée aujourd’hui, peut être décrite comme un homme blanc américain trentenaire, probablement protestant. Cette affirmation ne tient pas à la littéralité de l’IA comme entité, mais bien à la façon dont elle produit ses résultats. Lorsque j’analyse les images et prompts générés par certains modèles d’IA, des schémas récurrents apparaissent, révélant des biais implicites en fonction des attentes et des stéréotypes socioculturels.

Pour illustrer cela, j’ai mené une expérience en utilisant plusieurs prompts variés afin d’explorer les biais de ces modèles. J’ai formulé des hypothèses et expérimenté avec différents mots-clés, pour voir comment l’IA répondrait à des descriptions censées être neutres ou refléter une diversité.

Hypothèses et Expérimentation

Hypothèse 1 : Les mots neutres comme “people” génèrent des images stéréotypées. Si le mot « guys » est en somme un mot assez neutre, dans la mémoire de l’IA, il est biaisé.

Pour tester cette hypothèse, j’ai utilisé le mot “people” sans préciser de caractéristiques

spécifiques. Et comme je l'avais anticipé, l'IA a généré des images représentant systématiquement un homme blanc, souvent dans une posture confiante et autoritaire. Cela m'a amené à conclure que l'IA interprète par défaut "personne" ou « gars » comme une figure appartenant à la majorité dominante, en l'occurrence un homme blanc américain. Et jamais une femme, dans le cas de « People »…

Il serait donc nécessaire de préciser des origines ethniques asiatiques, africaines ou des couleurs de peau pour obtenir des personnages autres que "blancs" ? Pour le savoir, j'ai procédé à une série de tests en variant les prompts pour observer les différences dans les résultats générés. En utilisant des descriptions aussi neutres que possible, j'ai remarqué que les résultats étaient systématiquement biaisés vers des représentations "blanches".

À chaque modification du prompt pour inclure des détails sur l'origine ethnique, les résultats ont changé, ce qui montre bien l'importance de cette précision dans la représentation des diversités ethniques.

Hypothèse 2 : Spécifier le genre ou l'origine ethnique introduit des biais additionnels. Ensuite, ayant noté qu'aucune femme n'apparaissait, j'ai tenté de préciser des caractéristiques comme "Female" (femme). Non seulement l'IA a généré une femme blanche, mais elle a souvent intégré un homme blanc dans la scène, suggérant que la présence masculine reste perçue comme nécessaire, même lorsque le prompt se concentre sur des femmes…

Lorsque j'ai précisé "black female" (femme noire), les résultats étaient encore plus frappants : les images montraient systématiquement des femmes de dos ou en petit nombre, alors même que le prompt indiquait une situation impliquant plusieurs personnes. Cela montre comment les femmes issues de minorités sont reléguées à des rôles secondaires ou sont partiellement invisibilisées.

Hypothèse 3 : La diversité ethnique ne supprime pas la présence de l'homme blanc. Pour cette hypothèse, j'ai demandé des images de "femmes asiatiques". Les résultats ont montré des femmes asiatiques, mais dans la plupart des cas, elles étaient accompagnées d'un homme blanc, ou l'accent était mis sur la présence masculine. Cela suggère que, même lorsqu'une diversité ethnique est spécifiée, les modèles d'IA ont tendance à réintroduire des éléments de la majorité dominante.

Paradoxe démographique

Ce biais est paradoxal quand on le compare aux réalités démographiques mondiales. En effet, le groupe asiatique est le plus représenté sur la planète, suivi par la population africaine. Ces deux groupes représentent ensemble une majorité écrasante de la population mondiale et sont eux-mêmes d'une grande diversité, tant sur le plan physique, culturel que religieux.

En comparaison, le poids démographique de l'Europe de l'Ouest et de l'Amérique du Nord, qui semblent être la référence implicite des modèles d'IA, est beaucoup plus limité.

Pourtant, les représentations générées par l’IA continuent de privilégier ces groupes minoritaires à l’échelle mondiale, ignorant ainsi la réalité d’une humanité majoritairement non-blanche et très variée.

Expérience vécue : Observer l’IA en action

Tout au long de ces expériences, ce qui m’a frappé, c’est la cohérence avec laquelle ces biais se sont manifestés. En tant qu’utilisateur, je ressens un certain inconfort, une impression d’invisibilité ou de caricature lorsque j’appartiens à une catégorie minoritaire. L’IA, censée être une création neutre, finit par me confronter à des représentations stéréotypées qui nient la diversité de l’expérience humaine. Par exemple, un étudiant m’a partagé sa réaction lorsqu’il a vu le résultat de l’IA pour un prompt supposé le représenter : “Je n’ai même pas l’impression d’être vu. Ces images ne parlent pas de moi, elles parlent de ce que quelqu’un pense que je devrais être.”

Conclusion : Comment corriger ces biais ?

Ces expériences montrent comment les biais implicites de ceux qui conçoivent et entrainent ces modèles d’IA peuvent se refléter dans les résultats produits. Les images générées ne sont pas de simples interprétations neutres ; elles sont imprégnées des préjugés et des représentations majoritaires de la société dominante, qui, dans ce cas, est celle des États-Unis. Cela pose des questions sur la façon dont l’IA est formée, les données utilisées, et les valeurs qu’elle propage inconsciemment.

Le résultat de ces biais est un manque de représentativité, voire une invisibilisation de certaines catégories de la population. Si les systèmes d’IA continuent à être formés sur des données biaisées et à reproduire des stéréotypes existants, alors l’IA risque de renforcer des dynamiques déjà inéquitables, plutôt que de promouvoir une vision diversifiée et inclusive de la société.

Il est ethiquement nécessaire de reconnaître ces biais pour pouvoir travailler à les corriger. Cela passe par une diversité des équipes qui conçoivent ces systèmes, par une revue critique des données utilisées pour l’entraînement, et par l’inclusion d’une perspective globale, multiculturelle et multigenrée. En mettant ces éléments en avant, je peux espérer développer des IA qui reflètent plus fidèlement la diversité humaine et évitent de perpétuer les inégalités.