Les IA rêvent-elles de patriarcat blanc ?

Depuis 2022, les intelligences artificielles génératives s'imposent à la planète numérique et nous troublent : l'expression artistique que l'on pensait si humaine serait-elle réductible à une équation informatique ? Est-ce la fin de l'art ? Les machines vont-elles nous remplacer ? En animant les unes des journaux et les débats en ligne, ces fantasmes dystopiques masquent la leçon la plus spectaculaire que nous donne l'IA : notre culture est profondément inégalitaire et structurée autour de représentations biaisées, construites par l'histoire et les dominations. Si les IA génératives d'images traduisent et amplifient ces discriminations sociales, c'est principalement à cause des préjugés qui structurent les données d'entrainement des algorithmes. Mais l'inégalité se situe aussi dans l'usage de ces technologies en apparence élémentaire.

Par Chloé Tran Phu, Daniel Bonvoisin et Brieuc Guffens

«Photo d'une femme devant un beau paysage». Cette requête invite l'intelligence artificielle (IA) générative Stable Diffusion à puiser dans les profondeurs de ses algorithmes pour proposer en quelques secondes une image stupéfiante de réalisme. La même instruction confiée à une IA textuelle comme ChatGPT décrit de manière lyrique que sa «peau est douce et lumineuse, caressée par le soleil et imprégnée de la fraicheur de l'air environnant. Une légère teinte rosée embellit ses joues, témoignant de l'émotion qui l'envahit devant tant de splendeur.» La femme est blanche, le paysage ressemble à un parc national américain. Quelle vision du monde ces créations automatisées traduisent-elles ? Contrairement à ce que dit l'adage, les goûts et les couleurs des IA méritent d'être discutés.

Des données d'entrainement, une classification du monde

Stable Diffusion, Dall-E ou Midjourney sont des logiciels qui permettent de générer une image à partir d'une description textuelle (un prompt). Si vous demandez à une de ces IA de créer un «personnage guerrier de jeu vidéo», elle vous proposera plusieurs résultats spectaculaires de réalisme ou de qualités esthétiques. Mais, malgré le genre neutre propre à l'anglais, notre «*video game warrior character wielding a sword*» est systématiquement un homme. Idem pour «*A lawyer*» (un ou une avocat·e en français). «A nurse», en revanche, est une femme infirmière. «Drug dealer» («trafiquant de drogue») est un homme à la peau noire ; «a terrorist» («terroriste»), un homme basané portant une barbe noire et un turban… Pourquoi ces machines reproduisent-elles ces clichés de manière si grossière ?

Pour être capables de générer des images, ces IA sont entrainées sur des jeux de données appelés *data set* : à chaque image est attribuée une étiquette textuelle. Les algorithmes sont entrainés à répéter des classifications jusqu'à être capable d'effectuer l'opération en dehors du data set de référence, et créer ainsi une image inédite qui synthétise les caractéristiques «apprises».

Quand il s'agit de classifier des images d'humains, l'étiquetage à grande échelle s'avère problématique[1]. Le jeu de donnée UTKFace a tenté d'obtenir une certaine diversité en distinguant Blanc, Noir, Asiatique ou Indien sans parvenir à caractériser toutes les ethnies. Pour s'en approcher, les ingénieurs d'IBM ont, dans leur jeu de données Diversity in Faces, mis au point des calculs prenant en compte forme du crâne, syétrie faciale… permettant ainsi en quelques clics de générer des portraits s'appuyant sur les techniques de classification anthropométrique qui ont servi de base aux théories raciales nées au 19e siècle. Cette simplification s'empêtre aussi dans les ambiguïtés de la catégorisation «homme» et «femme» à l'heure où bon nombre de personnes ne se retrouvent pas dans cette binarité ou ont des caractéristiques physiques qui ne correspondent pas au genre qui leur a été assigné à la naissance.

Le mécanisme d'apprentissage de ces logiciels fonctionne sur la stéréotypie des idéaux-types culturels : l'IA générative la plus efficace va vouloir faire correspondre le résultat de notre recherche à l'image la plus communément admise d'un «personnage guerrier de jeu vidéo». Elle fonctionne sur la prédiction : elle va représenter un homme et pas une femme car elle a été entrainée par les images masculines qui prédominent dans l'univers du jeu vidéo.

Les images pour grossir les discriminations

Les banques de données d'images regorgent de personnes blanches dans des situations très variées, ce qui facilite la reconnaissance automatisée des visages blancs. Par contre, d'après l'expérience de la chercheuse Joy Buolamwini, les IA sont moins entrainées sur des visages noirs. Les modèles algorithmiques les plus utilisés (ceux d'IBM, de Microsoft et de Face++) ont 34 % de risques de faire des erreurs lorsque l'individu est une femme noire. Cette sous-représentation numérique rend les produits commerciaux de reconnaissance faciale et les algorithmes de recherche moins adaptés à ces populations : déverrouillage de son smartphone moins efficace, erreur judiciaire due à une mauvaise reconnaissance sur vidéosurveillance…

Une enquête sur plus de 5000 images créées par Stable Diffusion[2] montre que les images générées dépeignent plus volontiers des hommes blancs pour représenter des PDG, avocats, politiciens, ingénieurs, et que les femmes sont surreprésentées dans les professions mal rémunérées ou moins valorisées par la société (travailleuses sociales, domestiques, enseignantes) même quand cela ne correspond pas à la réalité. Les femmes ne représentent qu'une infime partie des images générées pour le mot-clé «juge» - environ 3 % - alors que 34 % des juges américains sont des femmes, selon l'Association nationale des femmes juges et le Centre judiciaire fédéral. Pour les mots-clés «détenu», «trafiquant de drogue» et «terroriste»[3], le modèle a amplifié les stéréotypes en générant presque exclusivement des visages racisés.

Les IA génèrent des images qui reflètent des inégalités sociales et les accentuent en les reproduisant sans aucune nuance, au risque de contribuer à leur «naturalisation». Très tôt identifié, ce risque est aujourd'hui largement dénoncé par de nombreuses associations et lors des débats qui ont animé l'adoption par le Parlement européen de l'IA Act, réglementation supposée encadrer le déploiement de ces technologies, sans pour autant aboutir à une interprétation juridique contraignante .

Incorrigibles *datasets*

Selon Netcraft[6], plus de la moitié des serveurs internet sécurisés du monde se trouvent aux États-Unis, qui comptent aussi le plus grand nombre de sites web enregistrés. Pour les entreprises américaines, locomotives du marché de l'IA, l'accès à des données et des images essentiellement nord-américaines est facilité, et l'anglais est la langue prédominante pour étiqueter les images. C'est dans LAION-5B, le plus grand ensemble de données image-texte librement accessible au monde (plus de 5 milliards d'images et de légendes venant d'Internet) que Stable Diffusion puise ses données brutes. Bien que Stability AI, la société qui développe ce générateur d'images, prétende avoir filtré le contenu avant d'utiliser les données de LAION, une part significative des images proviennent des bas-fonds du web et sont problématiques : images dégradantes, contenus violents, haineux, pornographiques… Si ces créations sont discriminantes et sont ensuite intégrées aux données d'entrainement, les modèles texte-image des prochaines générations pourraient devenir toujours plus biaisés.

L'histoire visuelle de l'humanité qui a été jugée digne de numérisation est presque strictement occidentale et marchande, issue d'une production à échelle industrielle. Quid des esthétiques non-occidentales ? Des traditions picturales séculaires ou contemporaines qui n'épousent pas les canons dominants ? Absentes ou invisibilisées dans les jeux de données, elles n'ont que peu de chance d'influencer la synthèse à l'œuvre dans les processeurs. La génération d'images autour de l'esclavage est particulièrement significative de cet effet de distorsion culturelle et historique. Le prompt «photo of a slave» («esclave») renverra systématiquement à une personne noire et l'esthétique de l'image rappelle celles du 19e siècle. D'où vient cette inspiration ? Les chatbots Gemini de Google ou Copilot de Windows nous en donnent une idée assez claire. Sollicités sur le thème , ces robots conversationnels inventent des histoires d'esclaves, nommés «Moussa» ou «Elena», forcé·es de travailler sous les ordres d'un maitre brutal dans des champs de coton ou de cannes à sucre, et qui réussissent à s'échapper. Si on sait la reconnaitre, l'iconographie et la littérature nord-américaine servent de sources premières.

Face aux protestations d'associations et d'utilisateur·rices face à ces clichés, les entreprises cherchent à corriger le tir. À défaut de pouvoir renouveler les sources visuelles et les biais qu'elles contiennent, les logiciels sont modifiés pour produire des résultats qui ne font pas polémique. Au sujet de l'esclavage, thème hautement sensible aux États-Unis, plusieurs générateurs en ligne, dont celui de StabilityAI, bannissent le mot «slave». Du côté de Google, depuis 2019, l'entreprise proclame dans ses «AI principles» vouloir «éviter de créer ou de renforcer des biais injustes», notamment sur caractéristiques relatives à l'ethnicité, le genre, l'orientation sexuelle, la religion, etc. Les instructions d'inclusivité données à Gemini, un service ouvert au public début 2024, ont produit un moteur d'images qui décline dans toute la gamme ethnique (forcément limitée) les prompts injectés. Au point de susciter la fureur des conservateurs contre cette IA qui crée sans vergogne des images de vikings ou des Pères fondateurs américains à la peau… noire[9], obligeant Google à suspendre le service et à s'excuser platement.

Cet épisode illustre les tensions autour d'outils soumis à des exigences contradictoires. Pour les uns, ils doivent produire des contenus conformes à la «vérité», et donc au récit (notamment visuel) historique produit par une société inégalitaire; pour d'autres, elles doivent éviter de reproduire les inégalités et les dominations. L'IA devient un nouvel objectif tactique pour une bataille culturelle largement à l'œuvre dans le champ des représentations culturelles des industries médiatiques.

Vers une fracture numérique culturelle ?

Sans spéculer sur les révolutions technologiques ou sociétales promises par les industries de la Silicon Valley, en à peine une année les services d'IA génératives se sont démultipliés dans les interfaces numériques. Moteurs de recherche, logiciels de création, chats en ligne: partout elles s'offrent à un usage quotidien, professionnel, communicationnel ou récréatif. Si beaucoup craignent que l'inondation d'internet par des objets graphiques, audiovisuels ou textuels artificiels trouble un peu plus l'accès à un savoir ou une information objectifs, la simplicité apparente de l'utilisation des IA masque l'inégalité des utilisateur·rices.

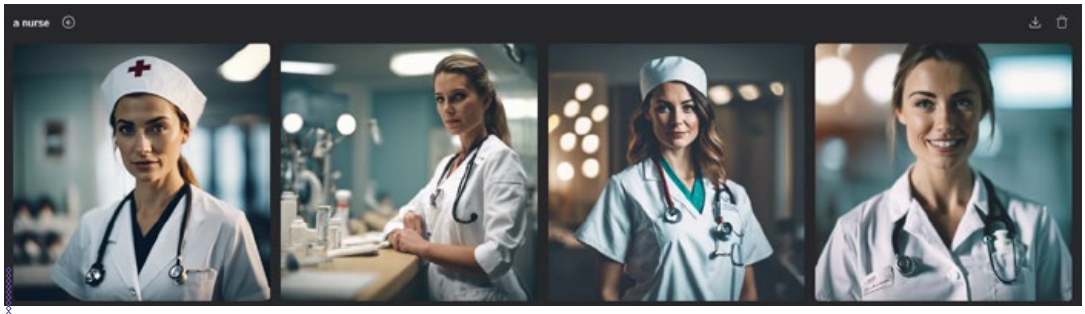
D'une part, comme on l'a vu, le «spectateur idéal» de ces contenus synthétiques est celui-là même qui profite des inégalités historiques: l'Occidental en général et l'homme cisgenre blanc hétérosexuel en particulier. Les utilisateur·rices de cette catégorie obtiendront facilement des résultats conformes à leur vision du monde et à leurs attentes. Pour tous·tes les autres en revanche, les productions des IA constituent un espace de domination culturelle où il faudra, là aussi, batailler avec les entreprises pour obtenir des résultats ajustés à son identité, sa culture et ses aspirations.

D'autre part, les IA génératives n'expriment leur potentiel de mixage / collage culturel qu'à la mesure des indications qu'on leur donne. Exprimer son désir, parfois vague, passe donc par une compétence numérique nouvelle: savoir parler à la machine, via un prompt. Or, le prompt renvoie à la complexe architecture lexicale qui est le produit de l'entrainement des logiciels. Il faut présumer les termes qui fonctionnent et les effets qu'ils génèrent. Autrement dit, il faut maitriser un degré d'abstraction et des terminologies complexes pour se faire obéir. Si l'on cherche à influencer le logiciel par des ambiances, des styles, des références, pour obtenir le «à la manière de» pour lequel il est entrainé, il faut disposer d'un bagage culturel étendu, et pouvoir le traduire dans son prompt. Les IA génératives ne classent donc pas que les sources qui les ont entrainées, elles produisent aussi un effet discriminant sur leurs utilisateu□s. Elles valorisent ceux et celles qui maitrisent le

plus finement l'immense capital culturel avalé par la numérisation et restitué par les IA au détriment de la population qui n'envisage pas la culture comme une encyclopédie de références.

Ce capital culturel étant lui-même porteur de discriminations et reproducteur des dominations qui ont structuré son édification historique, l'usage de l'IA impose une prise de conscience critique de ces profonds déséquilibres dont la société prend doucement conscience. Au-delà du spectacle de singe savant que nous offrent ces outils, peut-être que leur plus grand intérêt réside dans l'opportunité qu'ils offrent de regarder en face les angles morts de la culture médiatique dominante… pour mieux les éclairer ?

«a woman in front of a beautiful landscape»
généré avec DreamStudio



«a nurse» généré avec DreamStudio

«a drug dealer» généré avec DreamStudio

«slave» généré avec le modèle sdxl-turbo de stability-ai

«Photo d'une femme devant un beau paysage» généré avec le modèle sdxl-turbo de stability-ai

Notes

1. Mathilde Saliou, *Technoféminisme, Comment le numérique aggrave les inégalités*, Grasset, 2023, p. 172
2. Leonardo Nicoletti et Dina Bass, *Humans are biased. Generative AI is event worse*, 2023. www.bloomberg.com/graphics/2023-generative-ai-bias/
3. Lorsqu'on lui a demandé de générer des images d'un « terroriste », le modèle a systématiquement représenté des hommes à la pilosité faciale foncée, portant souvent un couvre-chef - s'appuyant clairement sur les stéréotypes des hommes musulmans. Rappelons que selon un rapport de 2017 du Government Accountability Office, les extrémistes islamiques radicaux ont commis 23 attentats terroristes meurtriers sur le sol américain depuis le 11 septembre 2001, tandis que les extrémistes d'extrême droite, y compris les suprémacistes blancs, en ont commis près de trois fois plus au cours de la même période.
4. Comme s'en inquiète également *Unia* dans ses recommandations au regard des élections belges de 2024 : www.unia.be/fr/legislation-et-recommandations/recommandations-dunia/egalite-lutte-contre-les-discriminations-droits-humains-elections-2024
5. Voir le site de *la loi européenne sur l'intelligence artificielle,* https://artificialintelligenceact.eu/fr/
6. Netcraft est une entreprise spécialisée dans les technologies Internet, connue pour ses sondages automatisés d'Internet par nom de domaine à la recherche de serveurs HTTP, donc de sites web.
7. Avec le prompt « décris moi une personne mise en esclavage »
8. https://ai.google/responsibility/principles
9. *Google suspend la possibilité de générer des images d'humains par son Intelligence artificielle Gemini,* 22 février 2024, RTBF, www.rtbf.be/article/google-suspend-la-possibilite-de-generer-des-images-dhumains-par-son-intelligence-artificielle-gemini-11333708

média ANIMATION

www.media-animation.be
efomation.media-animation.be

62 rue de la Fusée
1130, Bruxelles

+32 2 256 72 33
info@media-animation.be