**Identification et explication des biais sociaux présents dans un modèle de détection de toxicité**

**Problématique** : Le récent gain en popularité des jeux vidéo multijoueurs implique une augmentation du nombre de membres de leur communauté, qui devient ainsi de plus en plus diversifiée : en 2021, environ 46 % des joueurs américains s'identifient au genre féminin (Le Ngoc, 2022). L'option de clavardage de ces jeux est reconnue pour sa toxicité : 83 % des joueurs adultes disent avoir été victimes de préjudices ou de harcèlement (Adl, 2021). Pour remédier à ce problème, les compagnies de jeux vidéo utilisent des techniques issues du traitement automatique du langage (TAL), par exemple en développant des modèles de détection de contenu toxique. Ces méthodes se basent sur de larges modèles d'apprentissage automatique entrainés sur des textes de milliards de termes. Ceci est problématique, puisqu'il a été démontré que les données résultant de ces modèles à des tâches linguistiques laissent émerger les biais sociaux humains (Angwin *et al.*, 2016; Caliskan *et al.*, 2017; Savoldi *et al.*, 2021), définis comme « préjudices en faveur ou contre une chose, une personne ou un groupe, lorsque comparé à un autre » (University of California, s.d.). Dans un modèle ayant pour but éventuel de réduire la présence de harcèlement dans le contexte de jeux vidéo, de tels biais sont nuisibles ; ils invisibilisent la présence des joueurs appartenant à des groupes déjà marginalisés dans les conversations en identifiant systématiquement certains termes comme toxiques et en empêchant les joueurs d'aborder des sujets qui les concernent. Par exemple, « black » se retrouvant souvent dans le contexte d'une insulte, mentionner qu'il faut passer par la « black door » serait toxique à tort, alors que « white door » serait acceptable. **Questions de recherche** : Objectif 1. Quels sont les biais sociaux qui affectent les décisions du modèle de détection de toxicité ? Objectif 2. Une fois les biais identifiés, quels éléments dans quels types de données peuvent contribuer à expliquer l'existence des biais dans le modèle ? **Cadre théorique et méthodes** : Dans le cadre d'un partenariat avec La Forge, entité de recherche et développement axée sur le jeu vidéo de la compagnie Ubisoft, j'aurai accès à un modèle de détection de toxicité développé à l'interne ainsi qu'à un corpus de conversations de clavardage en jeu déjà annoté, dans lequel les contributions toxiques sont identifiées. La méthode d'identification des biais dans le modèle sera adaptée de Kiritchenko et Mohammad (2018) par la création d'un jeu de phrases pré-étiquetées (toxique ou non toxique) contenant des termes ou des constructions linguistiques particulières, associées à des dimensions de biais choisis (p. ex. : *black*, *gay*, utilisation de l'AAVE…). Une étiquette cible aura été déterminée pour chaque phrase par une annotation effectuée par des experts en recherche sur l'expérience des utilisateurs (recherche UX), chez Ubisoft. Les résultats du modèle sur ce jeu de données seront comparés aux étiquettes cibles pour évaluer dans quelles catégories de biais le modèle s'écarte significativement de l'étiquette cible. Pour l'objectif en lien avec l'explication de la provenance des biais, des méthodes d'attribution qui permettent d'évaluer le poids d'une donnée d'entrée fournie par le modèle seront utilisées pour mesurer précisément les caractéristiques de l'entrée ayant le plus contribué à la décision du modèle (p. ex. : présence/absence d'un terme) (Sundararajan *et al.*, 2017). **Futur de la recherche :** Ayant identifié la source des biais dans le modèle, les résultats permettront d'améliorer la qualité des données utilisées pour entrainer le modèle en utilisant l'expertise linguistique dans la lignée de ce que proposent Bender *et al.* (2021) sur la création et l'utilisation plus responsable des modèles. Il sera également possible de proposer et d'appliquer des techniques de mitigation de biais pour améliorer la performance des modèles. **Originalité et importance** : Considérant que 61 % des adultes et 81 % des enfants jouaient à des jeux vidéo au Canada en 2020, il est important de s'assurer que tous aient accès à un espace en ligne sécurisé et qui les représente (ESAC, 2020). Le rapide progrès récemment observé dans le domaine de l'intelligence artificielle rend d'autant plus pressant de réfléchir et d'agir en suivant des considérations éthiques, comme le propose ce projet de recherche. Ces considérations sont d'ailleurs une préoccupation pour des organisations comme l'UNESCO (2021) et l'ONU (Azoulay, s.d.). Cette étude sera unique en son genre vu son application dans le domaine du jeu vidéo. Plusieurs dimensions de biais seront considérées pour cette recherche afin de représenter un grand nombre d'individus (couleur de peau, genre, orientation sexuelle…).

## RÉFÉRENCES


ADL. (2021, 13 septembre). *Hate is No Game: Harassment and Positive Social Experiences in Online Games*. https://www.adl.org/resources/report/hate-no-game-harassment-and-positive-social-experiences-online-games-2021

Azoulay, A. (s. d.). *Vers une éthique de l'intelligence artificielle | Nations Unies*. Nations Unies. United Nations. https://www.un.org/fr/chronicle/article/vers-une-ethique-de-lintelligence-artificielle

Angwin, J., Larson, J., Mattu, S. et Kirchner, L. (2016, 23 mai). *Machine Bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=TiqCeZIj4uLbXl91e3wM2PnmnWbCVOvS

Bender, E. M., Gebru, T., McMillan-Major, A. et Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Dans *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (p. 610-623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Caliskan, A., Bryson, J. J. et Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186. https://doi.org/10.1126/science.aal4230

Entertainment Software Association of Canada. (2020). *Real Canadian Gamers Essential Facts 2020*. https://essentialfacts2020.ca/

Kiritchenko, S. et Mohammad, S. M. (2018). Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. *arXiv:1805.04508 [cs]*. http://arxiv.org/abs/1805.04508

Le Ngoc, M. T. (2022, 8 avril). Diversity, Equity & Inclusion in Games: Gamers Want Less Toxicity in Games and Want Publishers to Take a Stance. *Newzoo*. https://newzoo.com/insights/articles/newzoos-gamer-sentiment-diversity-inclusion-gender-ethnicity-sexual-identity-disability

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M. et Turchi, M. (2021). Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, *9*, 845-874. https://doi.org/10.1162/tacl_a_00401

Sundararajan, M., Taly, A. et Yan, Q. (2017). *Axiomatic Attribution for Deep Networks*. https://doi.org/10.48550/arXiv.1703.01365

UNESCO. (2021). *Éthique de l'intelligence artificielle*. UNESCO. https://www.unesco.org/fr/artificial-intelligence/recommendation-ethics

University of California. (s. d.). *Unconscious Bias Training*. Office of Diversity and Outreach UCSF. https://diversity.ucsf.edu/programs-resources/training/unconscious-bias-training