Bot Dog : une intelligence artificielle au secours de la justice sociale

Les intelligences artificielles ont récemment fait l'objet de vives critiques, notamment du fait de leurs biais discriminatoires et de leurs dérives déshumanisantes. Nous avons rencontré Morgane Bonvallat, cheffe du projet Stop Hate Speech de la Public Discourse Foundation, pour discuter du potentiel des IA pour une plus grande justice sociale.

Entretien : Danaé Leitenberg

Morgane Bonvallat, pourriez-vous nous décrire la genèse du projet Stop Hate Speech que vous dirigez et le Bot Dog développé dans ce cadre ?
L'idée de Stop Hate Speech est de lutter contre les discours de haine en ligne, en alliant la société civile, la technologie et la science. Notre premier projet dans ce cadre est la création d'une IA qui aiderait à contrer le discours haineux en ligne en Suisse. Cette idée a émergé en 2019, lorsque Sophie Achermann, directrice de la Public Discourse Foundation, a pris conscience de la problématique globale du discours haineux en ligne, notamment lors d'une session de l'ONU sur les violences faites aux femmes. Nous avons ensuite entrepris de créer un Bot Dog, c'est-à-dire un algorithme capable de repérer le discours haineux dans les sections de commentaire des sites internet de médias suisses et sur Twitter (avant son changement de nom et de propriétaire). Cet algorithme est le fruit d'une collaboration avec des chercheur·euse·s du Public Policy Lab de l'École Polytechnique Fédérale de Zurich (EPFZ) et du Digital Democracy Lab de l'Université de Zurich. Il se différencie de bon nombre d'algorithmes par sa qualité d'évaluation, car il doit son existence à un travail important, fourni par des humains : à l'aide d'une application, des utilisateur·ice·s en ligne, ainsi que des étudiant·e·s et des chercheur·euse·s de l'EPFZ ont pu évaluer plus de 40 000 commentaires postés sur des plateformes de médias suisses et Twitter en fonction de leur caractère haineux.

Que faut-il comprendre par discours haineux ?
Le discours haineux est un discours s'attaquant à un ou plusieurs facteurs identitaires de groupes d'individus, tels que le genre, l'âge, l'orientation sexuelle, l'appartenance religieuse, la nationalité, la couleur de peau ou l'origine. Cette définition est plus large que la définition proposée par l'article 261bis du Code pénal suisse, qui n'englobe pas les insultes sexistes par exemple.

Dans quel but cet outil a-t-il été développé et pourquoi avec les médias suisses en particulier ?
D'une part, il s'agit d'obtenir des données sur un phénomène dangereux et en pleine expansion, mais qui restait difficile à quantifier jusqu'ici. La Suisse représentait un cas intéressant parce que le discours haineux n'y avait pas fait l'objet d'études aussi systématiques. Si la problématique du discours haineux en ligne est globale, un projet national comme le nôtre fournit d'importantes informations et nous permet d'agir à échelle humaine. Une attention particulière a été portée aux subtilités culturelles, de langage et de contexte qui sont difficiles à repérer pour des algorithmes standards. D'autre part, il s'agit de visibiliser l'important travail de modération déjà fourni par les médias suisses. Les informations récoltées par le Bot Dog ont révélé qu'une grande partie des commentaires postés sur les plateformes des médias étaient modérés, c'est-à-dire approuvés (ou non) pour publication manuellement : cela signifie que des employé·e·s ont la difficile tâche d'évaluer du contenu

Morgane Bonvallat

« Internet est la plus importante extension de la vie publique que nous ayons connue. »

« Internet n'est pas un espace de liberté lorsqu'il n'est pas régulé. »

haineux en continu. En Romandie en particulier, certains médias ont fait le choix de fermer leurs sections de commentaires, en raison de la violence qui y florissait.

Quels sont les défis techniques que représente un tel outil ?

Un algorithme comme Bot Dog nécessite d'être entraîné constamment. Le langage utilisé et le contenu des discours haineux fluctuent au gré des événements mondiaux et mouvements sociaux. Le risque est donc de se retrouver avec un algorithme dont la précision est fortement diminuée.

Pourquoi le travail de modération fourni par cette IA est-il important ? Comment se répercute-t-il sur les questions d'inégalités de genre, d'appartenance, et de pouvoir au sens large ?

Une critique souvent formulée à l'encontre d'un projet tel que Stop Hate Speech est celle d'une soi-disant volonté de nuire à la liberté d'expression, car nous avons pour but de modérer les espaces virtuels et d'en éliminer les propos haineux. Mais cette critique omet une autre réalité : internet n'est pas un espace de liberté lorsqu'il n'est pas régulé. Nous savons que les individus responsables de publication de contenu haineux s'avèrent être un très petit pourcentage d'utilisateur·ice·s. Or, ces personnes prennent une place gigantesque dans le débat public. Nous ne pouvons pour le moment pas fournir de chiffres prouvant que les femmes* sont plus touchées par les propos violents en Suisse que les hommes cisgenres[1], mais nous savons qu'elles se comportent sur internet comme dans la rue : là où les hommes se sentent à l'aise, les femmes modifient leurs comportements pour éviter le danger, car elles ne s'y sentent pas en sécurité. Elles se retirent donc d'un pan crucial de la société. Comme le dit la journaliste Florence Hainaut, internet est la plus importante extension de la vie publique que nous ayons connue : nous consommons sur internet, nous travaillons sur internet et « relationnons » sur internet. Or, une large partie de la population y est réduite au silence, car maltraitée et exclue, lorsque cet espace n'est pas régulé. La liberté d'expression est-elle alors réellement garantie dans ces conditions ?

Concernant d'autres minorités, telles que les personnes racisées ou queers, elles voient leurs chances d'être harcelées multipliées, comme une étude le démontre avec le cas d'une femme noire dans l'espace virtuel.[2] Il s'agit donc d'une question de démocratie, une question morale et sociale qui se pose aussi en Suisse, où des politicien·ne·s et des activistes font l'objet de menaces de mort et de viol en raison de leur appartenance et/ou identité de genre. Certain·e·s, parfois représentant·e·s du peuple au niveau national, ont même fait le choix de ne plus s'exprimer sur certains sujets, tant le danger est grand.

Certain·e·s considèrent cependant qu'en répondant au discours haineux on l'attise. Qu'en pensez-vous ?

Simplement ignorer les commentaires violents en ligne ou limiter son utilisation des réseaux sociaux

pour éviter la violence n'est pas une solution viable. La vie virtuelle n'est pas séparée de la vie non-virtuelle, et la vie virtuelle n'est pas moins « réelle » que la vie non-virtuelle. Le harcèlement en ligne fait souffrir de manière très tangible et concrète, notamment les plus jeunes d'entre nous. La problématique des jeunes filles* victimes de cyberharcèlement est extrêmement inquiétante et peut avoir de graves conséquences, notamment sur la santé mentale et physique. Occulter ce problème, c'est considérer qu'il n'existe pas et ne pas s'atteler à le résoudre. Nous devons prendre cela au sérieux, pour le bien-être de ces jeunes générations en particulier, qui sont nées avec internet.

Avez-vous des recommandations d'actions à mener dans la lutte contre le discours haineux en ligne ?

Le Bot Dog nous fournit des données précieuses pour savoir à quoi l'on s'attaque, ce qui est un aspect primordial de la lutte contre le discours haineux en ligne. Nous développons actuellement un outil de visualisation qui indiquera les fluctuations du discours haineux en ligne en Suisse, en fonction de données récoltées auprès des médias. Cet outil sera un soutien pour la société civile que nous invitons à s'engager, notamment en prenant part à des actions de contre-discours. Une étude scientifique menée en collaboration avec l'EPFZ nous a permis d'observer que le contre-discours en ligne, par exemple en répondant à des commentaires haineux en exprimant de l'empathie avec le groupe affecté, est un moyen particulièrement efficace pour lutter contre le discours haineux.[3] Parallèlement, au sein de la Public Discourse Foundation, nous développons aussi un kit de premiers secours pour accompagner les individus ciblés par les violences. Nous souhaitons assister les personnes susceptibles d'être attaquées en ligne lors d'une intervention publique en les conseillant, les orientant, par exemple en mettant en place une communauté de soutien prenant en charge la signalisation des commentaires en ligne et le contre-discours. Nous pensons qu'il est également essentiel de lutter contre la violence virtuelle par la non-virtualité, en s'entourant de personnes de soutien, notamment pour les victimes de vagues d'agressions virtuelles.

Les IA sont souvent évoquées comme un symptôme négatif de notre temps, notamment à cause des utilisations qui en sont faites, souvent au détriment du bien-être humain. Quel regard portez-vous sur cette technologie après l'expérience Bot Dog ?

L'IA peut avoir des effets destructeurs, particulièrement lorsqu'il s'agit de la reproduction de stéréotypes de genre et/ou raciaux. L'idée derrière la création du Bot Dog était cependant d'utiliser cette technologie pour promouvoir plus d'égalité, de diversité et de démocratie. Pour garantir que l'intelligence artificielle soit au service de l'égalité, il est primordial que des femmes et des personnes issues de différents groupes marginalisés soient incluses dans la conception de ces nouveaux outils de communication. Nous nous trouvons dans un monde déjà façonné par l'IA et il est donc important de comprendre son fonctionnement si l'on veut la mettre au service d'une recherche de plus grande justice sociale. Notre projet a trouvé beaucoup d'échos dans les médias, mais aussi dans le monde politique et dans des offices fédéraux, car il répond aux besoins et préoccupations actuels de la société civile.

Danaé Leitenberg est docteure en anthropologie sociale et chercheuse à l'Institut de sociologie de l'Université de Bâle. Elle est spécialisée dans les questions liées aux inégalités sociales et à l'exclusion en Suisse ainsi que dans les conflits sociaux dans un contexte de modernisation croissante.

« Simplement ignorer les commentaires violents en ligne ou limiter son utilisation des réseaux sociaux pour éviter la violence n'est pas une solution viable. »

«Le harcèlement en ligne fait souffrir de manière très tangible et concrète.»

Notes

1 Un homme cisgenre est une personne qui s'identifie au genre masculin et dont l'identité de genre correspond au sexe assigné à la naissance.

2 Amnesty International : Toxic Twitter – A Toxic Place for Women, 2018. https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1-1/ (consulté le : 27.06.2024).

3 Hangartner, Dominik et al. : Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. In : Proceedings of the National Academy of Sciences 118 (50), 2021.

Abstracts

Bot Dog: eine KI im Dienst der sozialen Gerechtigkeit

In den letzten Jahren wurde Künstliche Intelligenz heftig kritisiert: Sie sei diskriminierend, sogar entmenschlichend, und widerspiegle oft die Auswüchse des technologischen Fortschritts. Es entstanden aber auch KI-Projekte, die die Gleichberechtigung und die soziale Gerechtigkeit fördern. Im Interview spricht **Morgane Bonvallat** über einen Algorithmus namens Bot Dog, der dafür trainiert wurde, Hatespeech in den Schweizer Onlinemedien zu erkennen. Das Projekt, das in Zusammenarbeit mit der ETH und der Universität Zürich durchgeführt wird, will in der Schweiz für Hatespeech im Internet sensibilisieren.

Bot Dog: un'IA al servizio della giustizia sociale

Negli ultimi anni, le intelligenze artificiali sono state aspramente criticate. Bollate come discriminatorie e persino disumanizzanti, sono spesso considerate emblematiche delle derive del progresso tecnologico. Parallelamente, hanno tuttavia visto la luce anche progetti che utilizzano le IA per promuovere l'uguaglianza e la giustizia sociale. Intervistata, **Morgane Bonvallat** presenta un algoritmo chiamato Bot Dog, addestrato per individuare i contenuti di odio online nei media svizzeri. Lanciato in collaborazione con il Politecnico federale e l'Università di Zurigo, il progetto intende sensibilizzare sui discorsi d'odio su Internet in Svizzera.