La Grande Conversation

ChatGTP et la démocratie

Société

Par Alberto Fernández Gibaja

Publié le 4 mai 2023

Parmi les premiers risques identifiés des outils de génération automatique de contenus, celui de la désinformation apparaît particulièrement préoccupant. Nous ne savons pas encore comment contrer la désinformation en ligne, qui prospère qui les réseaux sociaux. Il apparaît donc d'autant plus important d'anticiper les risques créés par la prolifération de contenus, potentiellement innombrables, générés par IA.

Cet article est tout d'abord paru sur le site d'Agenda Publica le 28 avril 2023.

Lorsqu'on demande à ChatGPT quels sont les dangers que son existence implique, l'une des réponses qu'il donne est la désinformation. Selon ChatGPT, et selon certains des plus grands experts mondiaux en la matière, la

prolifération des grands modèles de langage (LLM) crée une fenêtre d'opportunité pour un accroissement sans précédent de la désinformation. D'ores et déjà, alors que les contenus diffusés sur les réseaux sociaux sont créés par les utilisateurs humains, la désinformation en ligne affaiblit la démocratie. Il est temps de comprendre dans quelle mesure une croissance considérable de celle-ci grâce à l'IA générative peut accélérer cet affaiblissement et quels sont les outils à notre disposition pour l'en empêcher.

Il est important de rappeler en premier lieu ce que nous avons appris de la désinformation. Traditionnellement, on pensait que, comme le problème résidait dans le caractère mensonger du contenu – les fameuses « fake news » – alors le fait de s'attaquer au mensonge résoudrait le problème. D'où la prolifération du *factchecking* et la pression exercée sur les grands réseaux sociaux pour qu'ils mettent en place des systèmes de modération. Le *factchecking* comme la modération des contenus ont connu un succès pour le moins limité. Il est impossible de les mettre en œuvre à un niveau suffisant, comme le souligne Mike Masnick, dans son blog Techdirt. Modérer le contenu – décider ce qui peut ou ne peut pas être publié – sur des réseaux sociaux qui comptent des centaines de millions de messages par minute est tout simplement impossible et, comme il l'affirme, même si 99,9 % des messages étaient modérés, les 0,1 % qui ne le sont pas représenteraient des millions de messages.

C'est pourquoi de nombreux experts ont souligné que le problème n'est pas du côté du contenu, mais plutôt de celui des acteurs qui le diffusent et des techniques d'influence qu'ils utilisent.

Certains acteurs (partis politiques, candidats, gouvernements étrangers, groupes de pression, organisations sociales ou entreprises privées) cherchent parfois délibérément à manipuler et à tromper l'opinion. L'IA peut multiplier le nombre d'acteurs en abaissant les barrières à l'entrée. Les acteurs qui disposent

déjà des ressources nécessaires pourront multiplier de manière exponentielle la portée de leurs opérations d'influence. Dans le même temps, les sociétés qui se sont spécialisées dans ces services de désinformation peuvent devenir moins détectables et surtout beaucoup plus efficaces.

En termes de techniques d'influence coordonnées, l'IA peut améliorer et augmenter la portée des techniques existantes – telles que l'astroturfing, par exemple, ou la coordination multiplateforme – en générant plus de contenu, plus précisément, en moins de temps et à moindre coût. Par exemple, lors d'un événement tel qu'une manifestation de masse, il serait possible d'inonder les médias sociaux de faux contenus ou de contenus sans rapport avec la manifestation, ou de détourner l'attention, comme le fait déjà la Chine. Dans le même temps, de nouvelles techniques pourraient être créées, telles que des *chatbots* personnalisés qui tentent de persuader au niveau individuel.

Il est clair que la capacité de l'IA générative à produire de grandes quantités de contenus nuisibles et trompeurs est sans précédent. Dans les mois qui ont précédé le lancement de ChatGPT-4, OpenAI, l'organisation qui l'a créé, a monté une « équipe rouge » d'experts. Cette équipe a été chargée de tester le modèle de manière contradictoire, à la recherche de faiblesses potentielles pour générer des protections supplémentaires, et l'une de ses conclusions a été que l'utilisation du modèle pour produire du contenu trompeur est l'un de ses principaux dangers. Cela peut se produire à la fois dans des textes courts (par exemple, des messages sur les médias sociaux) et dans des textes plus longs tels que des rapports ou des discours politiques. Dans son livre blanc ChatGPT-4, OpenAI indique que ces modèles peuvent servir à renforcer les idéologies et les visions du monde, et rendre plus difficile la réflexion et l'amélioration de ces dernières. ChatGPT est un modèle qui ne produit – pour l'instant – que du texte, mais il existe d'autres outils qui produisent des images ou même

des vidéos, comme StableDiffusion, ce qui multipliera non seulement la quantité mais aussi la diversité des contenus nuisibles et trompeurs.

À cela s'ajoute le manque total de transparence de ChatGPT-4. OpenAI a décidé de ne pas publier les sources qu'elle a utilisées pour entraîner son modèle linguistique. Cela va à l'encontre de certains principes de ce que l'on appelle l'IA de confiance, notamment en ce qui concerne la protection de la vie privée, l'explicabilité – savoir pourquoi le modèle génère un résultat et pas un autre – et la transparence.

Comme on ne sait pas ce que le modèle utilise pour déduire ses résultats, il n'est pas possible de connaître les biais qui peuvent affecter ses productions. Ceci est particulièrement important parce que les grands modèles de langage sont des modèles généraux, c'est-à-dire qu'ils ne sont pas conçus pour une utilisation spécifique, mais peuvent être mis en œuvre pour une variété de services. Les acteurs politiques pourraient utiliser ChatGPT-4 pour concevoir des campagnes destinées à différents groupes sociaux sans connaître les biais possibles du modèle et, par conséquent, les biais qui apparaîtront dans ces campagnes.

Que pouvons-nous faire face à cette situation ? Il y a plusieurs domaines sur lesquels il est possible d'agir. D'une part, au niveau législatif, il faut lutter contre l'idée que toute réglementation est impossible parce qu'elle ralentirait la recherche, quand d'autres puissances ne le feraient pas, notamment la Chine. L'Union européenne est actuellement engagée dans la négociation de l'Artificial Intelligence Act. La législation, en cours d'élaboration depuis 2021, visait à réglementer les systèmes d'intelligence artificielle en fonction des dommages qu'ils pouvaient causer, en interdisant ceux qui pouvaient causer des dommages inacceptables et en créant des exigences strictes pour les autres. Cependant, des modèles tels que ChatGPT n'ont pas d'usage prédéfini et défient donc la réglementation avant même qu'elle ne soit adoptée. Ces

modèles constituent un casse-tête pour les agences nationales de supervision de l'intelligence artificielle, car la définition des risques est entravée par la généralité des usages.

A moyen terme, il faudrait envisager la création d'un système dans lequel les parties prenantes (États, organisations internationales, société civile, groupes de réflexion et entreprises privées, etc.) définiraient les principes directeurs de l'utilisation de l'intelligence artificielle. Ce système est en place depuis des décennies pour la gouvernance de l'infrastructure de l'internet et, bien qu'il ne soit pas parfait, il a réussi à établir un ensemble de principes et de règles qui ont permis à l'internet de rester relativement ouvert et libre. Certains experts ont même proposé de construire un système de démocratie délibérative qui soit garant des droits des personnes face aux systèmes d'intelligence artificielle.

La désinformation, ceux qui l'utilisent et les techniques qu'ils emploient, ne sont pas le problème, mais un symptôme d'une crise qui touche la démocratie au niveau mondial : l'affaiblissement progressif de la confiance dans les institutions qui devraient servir de contrepoids. Les médias, le pouvoir judiciaire, les forces de sécurité, mais aussi le Parlement, les diverses agences de régulation et une longue liste d'institutions devraient garantir l'intégrité des principes démocratiques. L'affaiblissement de ces institutions — et des droits qu'elles protègent — se produit à l'échelle mondiale, et la désinformation n'est qu'une arme parmi d'autres. ChatGPT et les modèles d'intelligence artificielle sont avant tout de nouveaux défis pour la démocratie.