

la majorité), le biais d'ancrage (qui pousse à se fier à l'information reçue en premier dans une prise de décision)...

L'emploi de *nudges* à travers des *bots* pour influencer les internautes s'est développé ces dernières années. Il peut s'agir de *nudges* pour inciter ou pour freiner des comportements, que Richard Thaler appelle *sludges* (littéralement «boues», en anglais). L'un des défis consiste à évaluer l'apport ou la menace qu'ils représentent. Ainsi, le projet *Bad Nudge-Bad Robot* que je dirige au CNRS à l'université Paris-Saclay vise à mettre en évidence le danger des techniques de *nudging* pour les personnes vulnérables telles que les enfants ou les personnes âgées. Ce projet entre dans le cadre de ma chaire d'intelligence artificielle Humaine (Human-Machine affective interaction & ethics) au CNRS, qui repose sur une forte collaboration interdisciplinaire entre l'informatique affective et la robotique émotionnelle, l'économie comportementale, la linguistique et le traitement du langage naturel.

DANS LA JUNGLE DU WEB

Par exemple, en 2019, dans une expérience de psychologie sociale conduite dans une école primaire, nous avons montré que les *nudges* des machines (robots, *chatbots*) étaient plus efficaces que les *nudges* d'humains. Dans cette expérience, la machine ou l'humain donne le même nombre de billes à chaque enfant. La première question posée à l'enfant est : combien il en garde pour lui et combien il en donne aux autres enfants. La deuxième question est un *nudge* pour le faire changer d'avis, en ajoutant par exemple que les autres en donnent plus (ce qui active le biais de conformisme). Plus les enfants étaient jeunes, plus ils étaient facilement manipulés par les machines. Cet exemple met en lumière le fort potentiel d'incitation associé à des agents capables de déployer un *nudge*. À n'en pas douter, de tels *bots* accroîtront les moyens de manipulation sur le web.

L'intelligence artificielle fournit ainsi des armes diverses, les unes permettant de traquer les propos haineux et repérer les *trolls* et les *bots*, les autres de propager des fausses informations et de manipuler les internautes, notamment grâce aux *nudges*. Ceux-ci sont aussi susceptibles d'être utilisés de façon positive. Imaginons par exemple qu'à la fin du mois on vous attribue des points en fonction de la qualité de ce que vous avez tweeté, et que vous puissiez comparer avec vos amis ou vos

collègues. Ce *nudge* vous inciterait à moins relayer des informations non vérifiées et peut-être à adopter des pratiques plus respectueuses des autres.

Lutter contre la désinformation s'annonce donc comme un défi permanent. Le recours à l'intelligence artificielle dans ce combat n'en est certainement qu'à ses débuts. Si se doter de tels outils sera utile pour faire face aux propos toxiques ou aux *bots* malveillants, cela restera néanmoins insuffisant.

Pour protéger citoyens et démocraties, une régulation de cette jungle qu'est le web, par des lois et des garde-fous éthiques, est nécessaire. C'est assurément une question aussi difficile que cruciale pour l'avenir de nos sociétés. Elle suscite aujourd'hui de nombreuses réflexions. Notamment, le Comité national pilote d'éthique du numérique a publié en juillet 2020 un document portant sur les enjeux dans la lutte contre la désinformation et la mésinformation, lequel identifie les tensions éthiques relatives à la mise en œuvre, par les plateformes, d'outils de modération de contenus et aux mécanismes de lutte contre la désinformation. Ce texte interroge aussi l'autorité acquise par ces plateformes et les rapports qu'elles entretiennent avec l'État et la justice. En résumé, il convient de redéfinir la responsabilité des plateformes aux niveaux national et européen, tout en protégeant la liberté d'expression. De ces sujets de société, nous allons devoir débattre.

— L'autrice —

> **Laurence Devillers**
est professeuse en intelligence artificielle à Sorbonne-Université et dirige une équipe de recherche au CNRS-LISN (Saclay), où elle est responsable de la chaire IA Humaine (<https://humaine-chaireia.fr>).

— À lire —

> **L. Devillers**, *Vague IA à l'Élysée*, L'Observatoire, 2022.

> **H. Ali Mehenni, L. Devillers et al.**, *Nudges with conversational agents and social robots: a first experiment with children at a primary school*, *Conversational Dialogue Systems for the Next Decade*, Springer, pp. 257-270, 2021.

> **C. Schneider et al.**, *Digital nudging: guiding online user choices through interface design*, *Communications of the ACM*, 2018.

> **D. Kahneman**, *Thinking, Fast and Slow*, Farrar, Straus and Giroux, 2011.

chantier en cours

C'est entendu, sous réserve d'être bien entraînés et gavés de données, les réseaux de neurones font des étincelles. Comment s'y prennent-ils ? C'est un problème : si l'intelligence artificielle symbolique était lisible, l'intelligence artificielle connexionniste, celle de l'apprentissage profond, est opaque. Parfois, la variation d'un seul pixel la fait dérailler. Beaucoup d'essais et d'erreurs sont nécessaires, car la théorie des réseaux neuronaux est très incomplète. Quant à savoir ce qu'ils comprennent... Pour la recherche, l'un des enjeux est donc de faire sortir cette branche de l'intelligence artificielle de son état de « boîte noire ».



Interro écrite pour l'intelligence artificielle

54



© Maggie Chiang