

En bref

> Sur internet, des outils d'intelligence artificielle sont utilisés de façon malveillante pour propager des informations fausses ou pour influencer les internautes.	> Ces outils exploitent nos biais cognitifs et nos réactions émotionnelles. Les <i>nudges</i> , qui consistent à inciter de façon imperceptible, renforcent leurs capacités de nuisance.	> Mais les <i>nudges</i> pourraient aussi aider les internautes à mieux résister à la désinformation.	> Des outils logiciels sont aussi développés pour traquer les contenus toxiques et repérer les robots malveillants œuvrant sur la Toile.
---	--	---	--

44

Avec la pandémie de Covid-19, nous avons vécu une période d'incertitude et d'anxiété qui a amplifié, en particulier sur internet, la propagation d'informations incomplètes ou fausses. Il y a eu d'abord la désinformation, c'est-à-dire la diffusion massive et délibérée d'informations visant à tromper ou à manipuler *via* les réseaux sociaux, les sites web, les forums, les messageries ou les *trolls* (internautes qui créent volontairement des polémiques). À laquelle s'est ajoutée la mésinformation, définie comme la propagation d'informations non vérifiées, incomplètes ou incertaines par des internautes qui n'ont bien souvent pas conscience des conséquences délétères de leurs actions.

Ainsi, compte tenu de l'arrivée massive d'informations provenant de sources multiples, contrôler la crédibilité et la fiabilité des données circulant en ligne devient une tâche urgente. D'autant que les techniques d'intelligence artificielle se développent et que les manipulateurs en tous genres n'hésitent pas à s'en servir.

Citons par exemple les techniques du *deep-fake*, qui permettent de truquer des vidéos en leur superposant d'autres bandes-son ou images. Par ailleurs, sur les réseaux sociaux, l'utilisation des *bots*, ou robots logiciels, devient de plus en plus fréquente. Il existe ainsi des comptes automatisés, gérés par un algorithme plutôt que par une personne réelle, qui peuvent influencer le comportement des internautes, émettre des avis et renforcer des opinions. Ces faux profils alimentent des réseaux complexes en administrant des groupes, en augmentant le nombre de leurs membres et en «likant» les messages sur les pages. On estime

que jusqu'à 15% des comptes Twitter actifs sont de faux comptes pilotés par des *bots*.

La menace va grandissant. Demain, ces *bots* seront capables d'interagir personnellement avec les internautes en détectant leurs émotions et leurs intentions. En exploitant nos biais cognitifs (*voir le schéma page ci-contre*), ils pourront nous manipuler de façon très efficace et nous inciter par exemple à adhérer à un parti politique ou à acheter des produits censés nous apporter du bien-être. Les défis portent donc, d'une part, sur la vérification de la crédibilité des informations émises ou relayées par les internautes et par les *bots*, et, d'autre part, sur la vérification des «intentions» de ces machines.

Les développements dans le domaine de l'intelligence artificielle devraient conduire à des outils de plus en plus efficaces pour traquer sur internet les propos haineux, offensants ou menaçants (*voir l'encadré pages 48 et 49*). Comme le préconisent la Commission européenne et le rapport 2019 de la mission française «Régulation des réseaux sociaux – Expérimentation Facebook», il est nécessaire de mener une réflexion d'ampleur en vue de constituer des bases de données communes pour améliorer les outils numériques de lutte contre la désinformation.

Des outils sont également disponibles pour détecter les faux profils et cartographier la diffusion de fausses nouvelles. Les ingénieurs de Facebook ont par exemple mis au point un outil d'apprentissage automatique qui a déjà permis de supprimer des milliards de faux comptes en 2020. Mais dès qu'un faux profil disparaît, un autre surgit. Pour mieux enrayer les dérives

Biais de confirmation

Tendance à accorder davantage de crédit et d'attention aux informations et événements qui confortent nos croyances, opinions ou préjugés.

Biais de négativité

Tendance à être davantage marqué par les expériences négatives que par les expériences positives.

Biais d'ancrage

Difficulté à se départir d'une première impression, qui pousse à se fier à l'information reçue en premier dans une prise de décision.

Effet Google

Tendance à oublier les informations qui sont facilement accessibles par une recherche sur internet, plutôt que de les mémoriser.

Biais de l'angle mort

Incapacité de l'individu à reconnaître ses biais cognitifs, alors qu'il les remarque chez autrui.

Illusion des séries

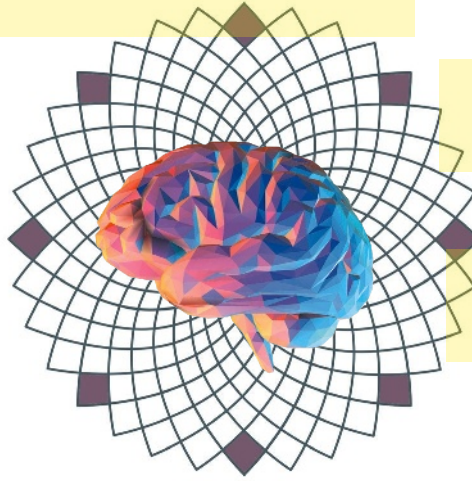
Tendance à trouver des régularités dans une série aléatoire d'événements.

Effet de l'influence continue

Tendance à croire une information reçue antérieurement même si elle s'est révélée fausse et qu'elle a été corrigée.

Biais d'autocomplaisance

Tendance de l'individu à attribuer ses succès à ses qualités propres et ses échecs à des causes externes, indépendantes de lui.



informationnelles sur les réseaux sociaux, il est donc nécessaire de comprendre les ressorts de la désinformation sur ces plateformes, mais aussi les manipulations possibles à partir de la détection de nos comportements.

De nouveaux terrains de recherche en intelligence artificielle s'ouvrent donc : identifier les fausses informations, les rumeurs et les préjugés, résoudre les conflits entre des sources de données hétérogènes, mais aussi détecter les *trolls* ou les *bots* qui cherchent à nuire en propageant des rumeurs ou en manipulant les opinions.

LES BIAIS COGNITIFS, LES ÉMOTIONS ET LE « NUDGE »

Sur les réseaux sociaux, la diffusion « virale », massive et rapide, de certaines informations tient en grande partie à la capacité d'attention limitée des internautes, à leur appétence pour les réactions émotionnelles et aux

↑ Notre raisonnement est souvent affecté par des biais cognitifs que les manipulateurs exploitent ou qui, par eux-mêmes, jouent un rôle essentiel dans la propagation virale des informations sur les réseaux sociaux. Les psychologues ont recensé plusieurs dizaines de biais cognitifs ; quelques-uns, parmi les plus répandus, sont indiqués ici.

biais cognitifs qui piègent nos raisonnements. Facebook, Twitter ou Instagram ont ainsi pour effet d'augmenter considérablement la visibilité des usagers malintentionnés ou de ceux qui sont prompts à exprimer sans frein leurs opinions. Certes, le principe de liberté d'expression est revendiqué par nos sociétés occidentales, mais les opinions exprimées sur les réseaux sociaux sont souvent des préjugés ou lieux communs dont la qualité ne devrait pas être estimée au nombre de clics qu'ils suscitent !

Les biais cognitifs jouent un rôle central dans la propagation de tels contenus. Et les futurs *bots* malveillants risquent de les exploiter de façon encore plus puissante et plus explicite. Dans les années 1970, les biais cognitifs ont été étudiés dans le cadre de l'économie comportementale, afin de comprendre nos décisions et agissements irrationnels. Les psychologues en ont recensé plusieurs dizaines, parfois classés en quatre catégories : les biais qui découlent d'un trop-plein d'informations, ceux dus à un manque de sens que l'individu cherche à combler, ceux liés au besoin d'agir vite et ceux induits par les limitations de la mémoire.

Le psychologue et économiste américano-israélien Daniel Kahneman, l'un des pères fondateurs de ce courant, Prix d'économie en mémoire d'Alfred Nobel en 2002, a montré dans ses travaux avec Amos Tversky le rôle important des émotions et des biais cognitifs dans nos processus de décision.

Une étape supplémentaire est venue avec un concept mis en exergue en 2008 par les Américains Richard Thaler et Cass Sunstein, et dont la mise en œuvre exploite les biais cognitifs et les réactions émotionnelles des humains. Ces deux chercheurs ont montré qu'à l'aide d'une conception appropriée des choix proposés aux individus, on peut améliorer de façon spectaculaire les décisions que ces individus prennent, sans forcer explicitement quiconque à faire quoi que ce soit. Richard Thaler et Cass Sunstein ont nommé *nudge* cette tactique de modification subtile du comportement.

Les *nudges* sont des sortes de coups de pouce imperceptibles qui prennent des formes très variées et qui sont utilisés dans divers contextes, notamment dans le marketing et la politique. L'exemple de *nudge* sans doute le plus connu est celui de la mouche peinte au fond de certains urinoirs : elle constitue une cible naturellement visée par les usagers, et aide ainsi à maintenir propres les toilettes.



Le *nudge* le plus célèbre est celui des urinoirs dont le fond est orné d'une image de mouche qui incite les usagers à viser cette cible et, ainsi, à maintenir la propreté des toilettes. Ce concept d'incitation/manipulation douce se répand aussi dans le monde numérique.