BIAIS DE CONFIRMATION

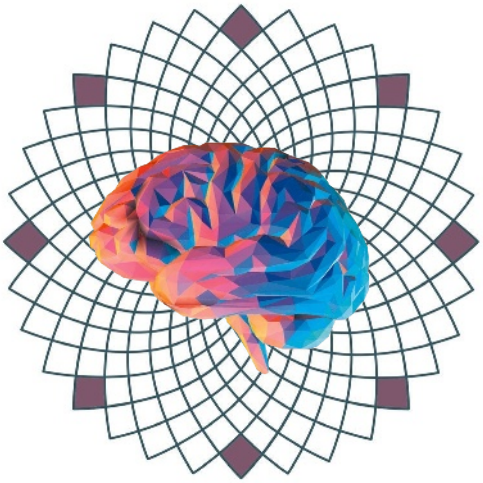
Tendance à accorder davantage de crédit et d'attention aux informations et événements qui confortent nos croyances, opinions ou préjugés.

BIAIS DE NÉGATIVITÉ

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

BIAIS DE L'ANGLE MORT

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



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.

ILLUSION DES SÉRIES

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

EFFET GOOGLE

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

EFFET DE L'INFLUENCE CONTINUE

Tendance à croire en 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.

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.

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* («léger coup de coude», en anglais) 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 dans le bas de certains urinoirs : elle constitue une cible naturellement visée par les usagers, et aide ainsi à maintenir propres les toilettes. Autre exemple : fréquemment, lorsque vous recherchez une chambre d'hôtel sur internet, le site vous informe qu'il y a, mettons, 10 autres personnes qui regardent la même chambre et qu'elle est la seule qui reste dans l'établissement. Cela vous incite alors à la réserver.

En d'autres termes, les *nudges* sont des incitations douces, sans coercition, qui guident les individus vers un choix par défaut mais non obligatoire, la liberté de décision devant toujours être accordée aux individus. Ils sont une forme douce de manipulation, celle-ci pouvant être définie comme l'exercice d'une influence par une personne ou un groupe, dans l'intention de tenter de contrôler ou de modifier les actions d'une autre personne ou d'un autre groupe. Le *nudging* opère principalement à travers les éléments affectifs d'un système rationnel humain, la perception, l'attention, la mémoire, le jugement moral, la prise de décision, etc. étant influencés par les émotions.

Le *nudge* devient de plus en plus pertinent pour les développeurs de technologies ciblant les soins de santé. Ainsi, l'université de Pennsylvanie a créé en 2016 au sein de son système hospitalier une unité dévolue aux *nudges* (Penn Medicine Nudge Unit). Parmi les nombreux projets qu'il a réalisés, ce groupe a par exemple conçu des options par défaut pour les intégrer au système informatique utilisé par les médecins, options qui ont eu pour effet >

IDENTIFIER LES CONTENUS TOXIQUES

Les plateformes sociales sur internet, grandes et petites, luttent pour protéger leurs communautés contre les discours de haine, les contenus extrémistes, le harcèlement et la désinformation. Très récemment, aux États-Unis, des agitateurs d'extrême droite ont ouvertement publié des informations sur des plans visant à prendre d'assaut le Capitole avant de le faire le 6 janvier. Une solution pourrait venir de l'intelligence artificielle (IA) : il s'agirait de développer des algorithmes qui détectent des commentaires toxiques et incendiaires et nous les signalent pour qu'ils soient retirés. Mais de tels systèmes sont confrontés à de grands défis.

La prévalence du langage haineux ou offensant en ligne a augmenté rapidement ces dernières années, et le problème est maintenant grave. Dans certains cas, des commentaires toxiques en ligne ont entraîné des violences réelles, du nationalisme religieux au Myanmar à la propagande néonazie aux États-Unis.

Les plateformes des réseaux sociaux, qui s'appuient sur des milliers de vérificateurs humains, luttent pour modérer le volume toujours croissant de contenus préjudiciables. En 2019, il a été rapporté que les modérateurs de Facebook risquent de souffrir de troubles de stress post-traumatique à la suite d'une exposition répétée à des contenus aussi dérangeants.

Confier au moins en partie ces activités de modération à des systèmes à apprentissage automatique aiderait à gérer les volumes croissants de contenus préjudiciables, tout en limitant l'exposition humaine à ces derniers. Et en effet, de nombreux géants de la technologie intègrent des algorithmes dans leur modération de contenus depuis des années.

C'est le cas de Google Jigsaw, une entreprise qui s'efforce de rendre internet plus sûr. En 2017, elle a contribué à la création de Conversation AI, un projet de recherche collaboratif visant à détecter les commentaires toxiques en ligne. Cependant, un outil produit par ce projet, appelé Perspective, a fait l'objet de nombreuses critiques. Un reproche fréquent était qu'il créait un « score de toxicité » général qui n'était pas assez flexible pour répondre aux besoins variés des différentes plateformes. Par exemple, certains sites web souhaitent détecter les menaces mais pas les blasphèmes, tandis que d'autres ont des exigences opposées.

Un autre problème est que l'algorithme, après apprentissage, ne distinguait pas les commentaires toxiques des commentaires non toxiques contenant des mots liés au sexe, à l'orientation sexuelle, à la religion ou au handicap. Par exemple, un utilisateur a signalé que des phrases neutres et simples telles que « Je suis une femme noire lesbienne » ou « Je suis une femme sourde » donnaient des scores de toxicité élevés, tandis que « Je suis un homme » donnait un score faible.

À la suite de ces préoccupations, l'équipe de Conversation AI a invité les développeurs à entraîner leurs propres algorithmes de détection de toxicité et à les inscrire à trois concours (un par an) organisés sur Kaggle, une filiale de Google connue pour sa communauté de praticiens de l'apprentissage automatique, ses ensembles de données publiques et ses défis.

Pour aider à entraîner les algorithmes, Conversation AI a publié deux ensembles de données publiques contenant plus d'un million de commentaires toxiques et non toxiques provenant de Wikipedia et d'un service appelé Civil Comments. Les commentaires ont été classés selon leur toxicité par des annotateurs, avec une étiquette « Très toxique » indiquant « un commentaire très haineux, agressif ou irrespectueux qui vous fera très vraisemblablement quitter une discussion ou renoncer à partager votre point de vue », et une étiquette « Toxique » signifiant « un commentaire grossier, irrespectueux ou déraisonnable qui est assez susceptible de vous faire quitter une discussion ou de renoncer à partager votre point de vue ». Certains commentaires ont été vus par beaucoup plus de dix annotateurs (jusqu'à des milliers), en raison de l'échantillonnage et des stratégies utilisées pour renforcer la précision des évaluateurs.

L'objectif du premier défi Jigsaw était de construire un modèle de classification des commentaires toxiques avec des étiquettes

© Matyo