

> 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 (IA) se développent et que les manipulateurs en tous genres n'hésitent pas à s'en servir.

Citons par exemple les techniques du *deepfake*, qui permettent de truquer des vidéos en leur superposant d'autres bandes-son ou images. Par ailleurs, sur les réseaux sociaux, l'utilisation de «bots», des 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 la figure 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 30 et 31). Comme le préconisent la Commission européenne et le rapport (2019) de la mission « Régulation des réseaux sociaux - Expérimentation Facebook » commanditée par le gouvernement français, il est nécessaire de mener une réflexion d'ampleur sur la constitution de bases de données communes pour améliorer les outils numériques de lutte contre la désinformation et la mésinformation.

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 est à terre, un autre surgit. Pour mieux enrayer

les dérives informationnelles sur les réseaux sociaux, il est donc nécessaire de comprendre les ressorts de la désinformation sur ces plateformes (voir l'article pages 34 à 41), mais aussi les manipulations possibles à partir de la détection de nos comportements.

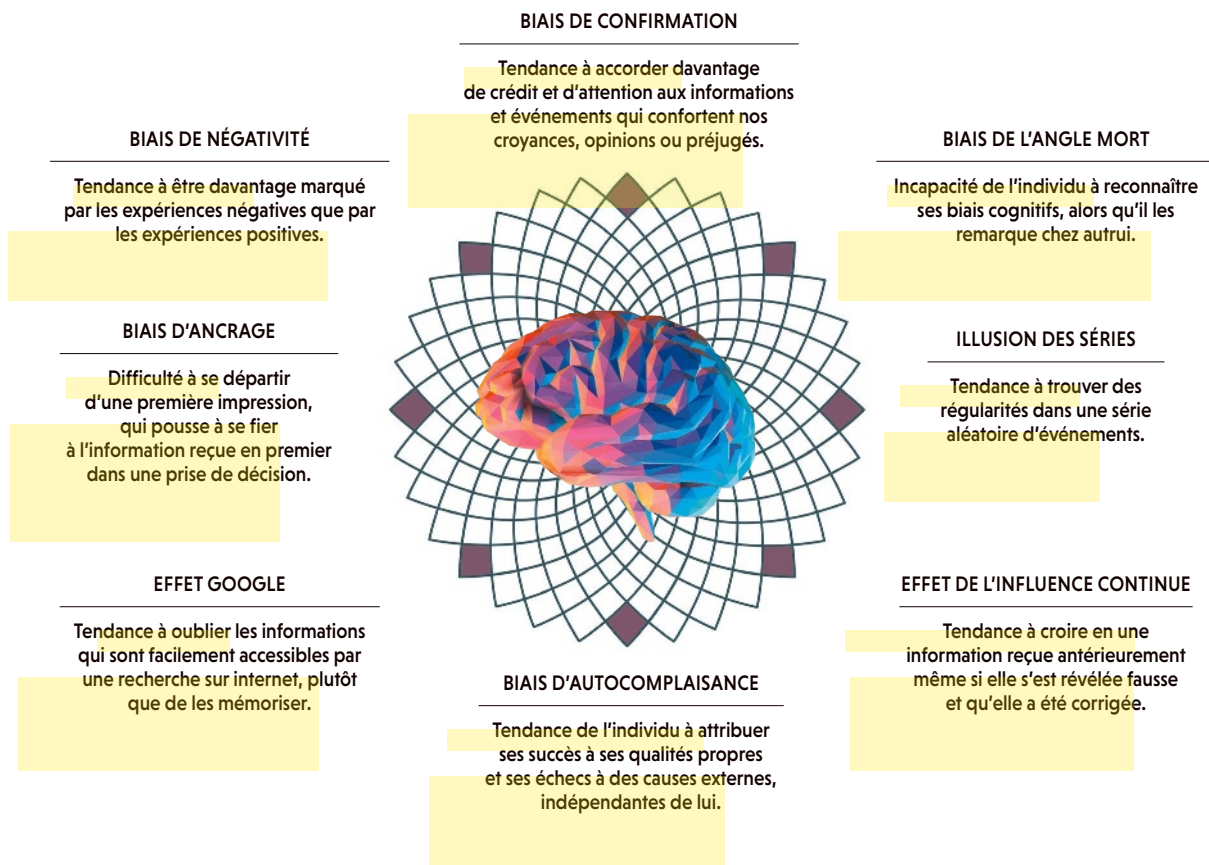
De nouveaux terrains de recherche en intelligence artificielle s'ouvrent donc : identifier les fausses informations, les rumeurs et les préjugés, résoudre les conflits entre des sources de données hétérogènes, mais aussi détecter les trolls ou les bots qui cherchent à nuire en propageant des rumeurs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

LES BIAIS COGNITIFS, LES ÉMOTIONS ET LE « NUDGE »

Sur les réseaux sociaux en ligne, la diffusion «virale», c'est-à-dire de façon 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 biais cognitifs qui piègent nos raisonnements. Les plateformes numériques telles que Facebook, Twitter ou Instagram ont ainsi pour effet d'augmenter considérablement la visibilité des usagers malintentionnés ou de ceux qui sont prompts à exprimer librement des 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 jugements rapides, des idées reçues, des évidences, des préjugés, des lieux communs... Des contenus dont la qualité ne se juge pas au nombre de clics qu'ils suscitent !

Les biais cognitifs jouent un rôle central dans la propagation de tels contenus et de la désinformation. 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 comportements 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

En 2020, Facebook a supprimé, grâce à un outil d'IA, des milliards de faux comptes



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

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.

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 >