**Forces et faiblesses de l’IA dans la détection de fake-news**

**Forces**

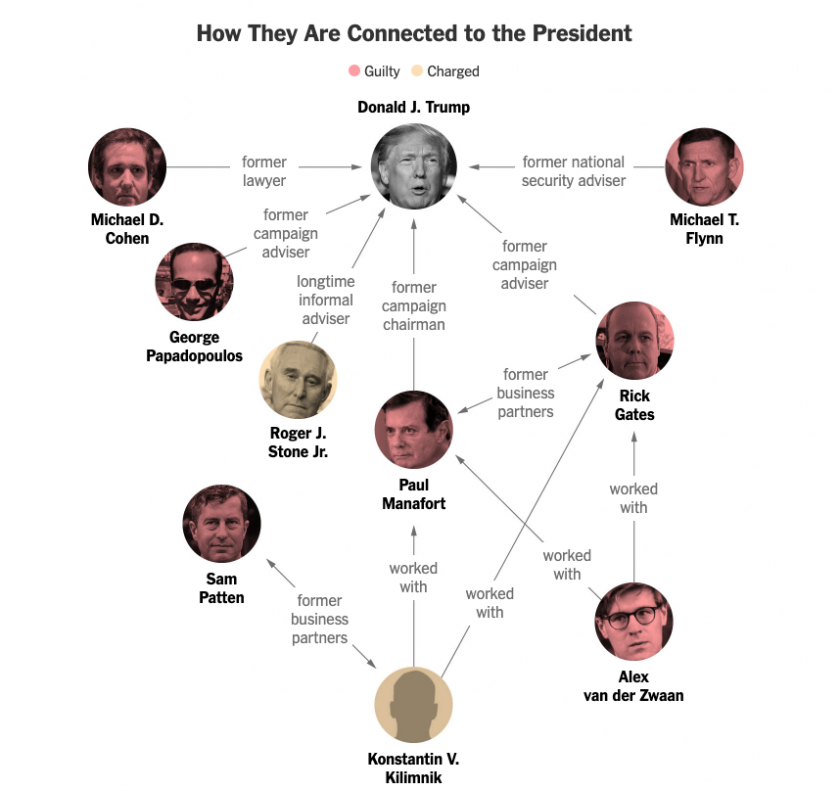
* De nos jours, des techniques d’intelligence artificielle peuvent aider à repérer les textes similaires, en tenant compte du contenu, mais aussi du canal de diffusion, de la personne qui relaie le message, et d’autres éléments contextuels, par exemple les images et illustrations. On est alors proche du mode de fonctionnement des moteurs de recherche : les modèles de recherche d’information actuels sont plutôt efficaces pour retrouver des textes similaires, même s’ils n’emploient pas exactement les mêmes mots ou les mêmes tournures :

Mais le but ultime serait évidemment de repérer directement les fake news par des moyens automatiques. Ceci semble en fait extrêmement difficile en l’état des choses et, si on écarte les faux grossiers, même un humain aura du mal à caractériser certains textes. Plusieurs techniques sont explorées en intelligence artificielle.

* La première technique consiste à repérer des informations factuellement fausses en comparant un texte donné avec les informations contenues dans une base de données. Ceci peut fonctionner en théorie (un jeu de donné appelé FEVER, pour *fact extraction and verification*, a même été développé pour cela), mais on dispose rarement de bases de connaissances adaptées au problème. En gros, l’actualité ne se réduit pas à une base de données et les fake news ne portent pas tellement sur des informations factuelles isolées.
* Une seconde technique est de repérer des documents types, grâce à leur titre, leur mise en page, les illustrations qui les accompagnent, entre autres. Ceci peut aussi fonctionner jusqu’à un certain point, mais ce n’est pas très précis. Par exemple, de nombreux titres racoleurs utilisent une mise en page tape-à-l’œil sans pour autant être des fake news.
* L’IA en faveur de l’investigation ?

Il est important de sortir d’un prisme déformant qui présente l’IA comme source de troubles et qui vole les compétences des journalistes. Souvenons-nous du film Spotlight (Tom McCarthy, 2015). Une poignée de journalistes enquête alors sur la culpabilité de l’Église catholique dans des scandales de pédophilie. Le journaliste Matt Carroll réussit, précisément grâce à la data, à croiser les différentes affaires. Mais depuis 2002, les outils technologiques se sont transformés : « L’augmentation toujours grandissante de la puissance de calcul et de la performance des algorithmes va permettre de collecter et de traiter encore plus rapidement des données de plus grande taille », synthétise le rapport.

Et nous en sommes déjà là, avec l’émergence du data journalisme que le rapport considère comme étant « à l’origine d’un nouveau journalisme d’investigation ». Une approche où l’homme et la machine collaborent au lieu de s’opposer et qui permet au journaliste de se concentrer « sur des interviews et des échanges privilégiés avec des individus mieux identifiés ». Ce qui pourrait d’ailleurs être une réponse intelligente à la défiance envers les médias, qu’on accuse souvent de tomber dans le piège de l’immédiateté de l’information et de la course aux clics.



**Faiblesses**

La première limite identifiée est structurelle : la solution mise en avant est précisément la cause du problème. Le *deepfake* est un parfait exemple de la menace informationnelle permise par l’intelligence artificielle. Cette technique permet la création et la modification de contenu audio-visuel d’une qualité telle qu’il est très complexe d’identifier la manipulation. Les images et vidéos, auparavant mise en avant comme preuves irréfutables ne pourront plus être considérées comme telles. Par ailleurs, les cybers criminels ont d’ores et déjà intégré l’intelligence artificielle à l’attirail d’outils qu’ils utilisent. Déjà en 2017, 62% des conférenciers, experts de la cybersécurité, du *Black Hat* de Las Vegas estimait déjà que les cybers criminels utiliseraient l’intelligence artificielle pour passer à l’offensive.

On dit souvent de l’IA (et des nouvelles technologies en général) qu’elle est une épée à double tranchant. D’une part, elle permet l’émergence de menaces informationnelles en ligne de plus en plus sophistiquées et l’abaissement des remparts contre les acteurs malveillants. L’IA serait notamment utilisée pour cadrer plus précisément les paramètres d’une attaque informationnelle : quoi, qui et quand attaquer. D’autre part, l’IA présente de nombreuses opportunités pour la lutte contre ces mêmes menaces informationnelles.

* La seconde limite est la capacité de l’IA à se tromper. L’intelligence artificielle est aujourd’hui très au point sur la détection sémantique. Seulement certains aspects de langage lui sont encore complexes à appréhender. Il s’agit par exemple du sarcasme, de la persuasion ou de l’ambivalence. C’est une des raisons qui explique l’existence de *faux positifs,* limites majeures de l’IA. Le fait de désigner un post comme étant de la désinformation, un abus ou un troll alors qu’il ne le serait pas, minimise la propagation des outils de détection et leur efficacité.

Aussi, pour une analyse efficace, l’IA a besoin de jeux de données mais n’en dispose pas toujours pour chaque problème qui se pose (problèmes de Copyright notamment ).

* Il faudrait ajouter que les plates-formes ont alors beau jeu d’en appeler à la régulation de la part des États. Mark Zuckerberg a ainsi dit : « nous ne souhaitons pas que les entreprises privées prennent autant de décisions importantes qui touchent aux valeurs fondamentales sans contrôle démocratique » (« We don’t want private companies making so many decision – balancing social equities without democratic processes »), tout en protestant quand une nouvelle législation se met en place.
* La dernière limite réside dans l’action humaine en amont de l’utilisation de la machine. Les outils utilisant le Machine-Learning sont paramétrés par des Hommes, qui ont eux-mêmes leurs propres biais cognitifs. Se pose ainsi la question de l’objectivité du paramétrage de la machine pour déterminer ce qu’est une fausse information et ce qui n’en est pas. Certaines plateformes utilisent même l’IA pour identifier des contenus légaux mais définis par ces mêmes plateformes comme étant nuisibles : « comportements inauthentiques » et « post insensibles » sont par exemple visés. Au-delà de la question de l’objectivité, comment admettre qu’une plateforme définisse la limite entre le sensible et l’insensible, l’authentique et l’inauthentique.