PREMIER MINISTRE

Liberté
Égalité
Fraternité

Secrétariat général de la défense et de la sécurité nationale

Défis et opportunités de l'intelligence artificielle dans la **lutte contre les manipulations de l'information**

VIGINUM

Enjeux systémiques

SOMMET POUR L'ACTION SUR L'IA

Le présent rapport a été réalisé par VIGINUM avec les contributions internationales issues, pour les acteurs institutionnels, du Service Européen pour l'Action Extérieure (SEAE), de l'Agence de Défense Psychologique (Suède), du Mécanisme de Réponse Rapide du Canada (MRR Canada), du bureau des Affaires étrangères et du Commonwealth (Royaume-Uni), et pour la société civile, de deux organisations indépendantes de *fact-checking* : *Lupa* (Brésil) et *Full Fact* (Royaume-Uni).

Au titre des attributions qui lui sont confiées par l'article 3 du décret n°2021-922 du 13 juillet 2021, le service de vigilance et de protection contre les ingérences numériques étrangères (VIGINUM) a pour missions de détecter et de caractériser les opérations d'ingérences numériques étrangères (OINE) en analysant les contenus publiquement accessibles sur les plateformes en ligne. Pour ce faire, VIGINUM est autorisé par le décret n°2021-1587 du 7 décembre 2021 à opérer un traitement automatisé de données à caractère personnel.

Volet numérique de la manipulation de l'information, l'ingérence numérique étrangère est une opération « *impliquant, de manière directe ou indirecte, un Etat étranger ou une entité non étatique étrangère, et visant à la diffusion artificielle ou automatisée, massive et délibérée, par le biais d'un service de communication au public en ligne, d'allégations ou imputations de faits manifestement inexactes ou trompeuses de nature à porter atteinte aux intérêts fondamentaux de la Nation* ».

Depuis le début des années 2020, les applications d'intelligence artificielle (IA), notamment générative (IAg), ont connu un développement sans précédent. L'ergonomie des produits ainsi que leur accès facile et gratuit ont en effet favorisé leur appropriation par le plus grand nombre : ainsi, selon BPI France[1], d'ici 2027, le monde pourrait compter un demi-milliard d'utilisateurs réguliers de technologies liées à l'IA.

Si certains développements laissent entrevoir des perspectives prometteuses en matière de santé, d'environnement ou de mobilité, la diffusion massive des outils génératifs interroge néanmoins quant à ses conséquences sur l'intégrité de l'environnement informationnel. En particulier, elle fait craindre une élévation structurelle du niveau de menace liée aux ingérences numériques étrangères, avec le risque d'une altération de la perception de la réalité par les citoyens. En effet, avec la généralisation des outils d'IA, et par conséquent, la probable profusion de contenus générés artificiellement sur les plateformes en ligne, il deviendra plus difficile de distinguer l'authentique du synthétique. En matière informationnelle, les conséquences d'une réalité travestie sont susceptibles de produire des effets durables et négatifs sur le fonctionnement de la démocratie, notamment lors de rendez-vous électoraux, et plus largement conduire à la déstabilisation de nos sociétés, fondées sur la notion de confiance.

Aussi, dans un contexte international marqué par le rapport de force et les interdépendances, dans lequel la maîtrise de la technologie s'affirme comme un facteur de puissance majeur, il est utile de s'interroger sur les conséquences réelles de l'IA sur la menace informationnelle, tout en explorant également ses cas d'usage au service de la lutte contre les manipulations de l'information.

Au-delà des promesses et des peurs liées à l'usage de l'IA, ce rapport se donne ainsi pour ambition d'appréhender ces enjeux technologiques de manière réaliste : si les technologies d'IAg sont susceptibles d'accroître certaines capacités de la menace informationnelle, l'IA offre également de réelles possibilités de renforcer nos défenses face aux manœuvres informationnelles, et ce quel que soit leur degré de sophistication.

Présentant les points de vue croisés de VIGINUM et d'acteurs internationaux, à la fois institutionnels et issus de la société civile, ce rapport entend répondre à trois objectifs principaux : **contribuer à l'information et à la sensibilisation du public,** en élevant le niveau de connaissance des cas d'usage malveillants de l'IA à des fins de manipulations de l'information ; **mettre en évidence les opportunités offertes par l'IA** pour la lutte contre les manipulations de l'information, afin de promouvoir le partage international de bonnes pratiques ; et enfin, **encourager la coopération** entre les acteurs institutionnels, la société civile, le monde académique et scientifique et les acteurs privés, afin d'accélérer le développement de solutions innovantes au bénéfice de l'ensemble de l'écosystème en charge de la lutte contre les manipulations de l'information.

[1] Source : BPI France le Hub, sur la base de données Statista, cf https://bigmedia.bpifrance.fr/nos-actualites/marche-de-lintelligence-artificielle-ou-en-sommes-nous ; et https://comarketing-news.fr/le-marche-de-lia-va-doubler-en-4-ans/

I. L'IA, une technologie utilisée par les acteurs de la menace informationnelle

1. Définitions

L'intelligence artificielle (IA) désigne l'ensemble des procédés logiques et automatisés, reposant généralement sur des algorithmes, destinés à reproduire, au moins partiellement, des comportements humains, tels que l'apprentissage, le raisonnement, la planification ou la création.

L'intelligence artificielle repose sur trois grands concepts :

- L'apprentissage automatique, ou *machine learning,* vise à conférer aux machines la capacité d'apprendre de manière autonome, *via* des modèles mathématiques, afin d'extraire les informations les plus pertinentes d'un ensemble de données d'entraînement mis à leur disposition ;
- Sous-catégorie du *machine learning*, l'apprentissage profond ou *deep learning* est un procédé d'apprentissage automatique utilisant des réseaux de neurones artificiels permettant la résolution de problèmes complexes tels que la reconnaissance des formes (vision artificielle) ou le traitement du langage naturel ;
- Enfin, sous-segment applicatif du *deep learning* et permettant de générer du texte, des images ou encore du code informatique, les technologies d'IA générative (IAg) constituent un changement d'échelle dans la génération de contenus. Ces technologies ont réalisé une percée majeure en 2022, avec la mise à disposition pour le grand public de services *web* permettant de générer très facilement des contenus textuels et des images.

2. Une technologie attractive pour les acteurs de la menace informationnelle

Nota bene : Les tendances et cas d'usages documentés dans les parties 2) et 3) couvrent uniquement les modes opératoires utilisant l'IA générative, observés dans le cadre d'investigations en sources ouvertes. Les usages potentiels d'autres segments d'IA dans le cadre de manœuvres informationnelles (usage d'algorithmes de machine learning *par exemple, pour le ciblage d'audiences spécifiques) ne sont ici pas couverts, ceux-ci n'étant pas aisément documentables en sources ouvertes.*

Dans le contexte d'une utilisation désormais largement généralisée des technologies d'IAg dans de nombreux domaines, VIGINUM observe depuis deux ans un recours croissant à l'IAg par les différents acteurs étrangers de la menace informationnelle.

Trois tendances principales caractérisent les conséquences de cette technologie sur la physionomie de la menace informationnelle actuelle :

- Un changement d'échelle dans la génération de contenus potentiellement inexacts ou trompeurs ;
- Des capacités décuplées pour la réplication et la publication coordonnée de contenus inauthentiques à grande échelle ;
- Une aide pour la génération et la gestion de comptes inauthentiques sur les plateformes en ligne.

a. Un changement d'échelle dans la génération de contenus potentiellement inexacts ou trompeurs

A l'instar des capacités nouvelles offertes par l'IA pour de nombreux secteurs d'activité (santé, transport, environnement), l'usage de technologies d'IAg est susceptible d'accroître significativement les capacités

des acteurs malveillants à générer des contenus manifestement inexacts ou trompeurs dans le champ informationnel, *via* majoritairement trois type d'usage.

Tout d'abord, ces technologies permettent **l'optimisation de leur productivité basique** (traduction, correction grammaticales et orthographiques, synthèse de documents…).

Par ailleurs, elles permettent **la génération à grande échelle de textes sous différentes formes**. Il peut s'agir tout à la fois de textes longs (articles par exemple), ou courts (publications pour les réseaux sociaux, commentaires…), ou encore de textes pouvant viser un public large, ou bien de textes sur mesure ciblant des audiences très spécifiques (*via* des *prompts*[2] visant à imiter la tonalité et le langage de *personas* représentatifs de certaines catégories de population à des fins de *microtargeting*). Cet usage recouvre des niveaux de sophistication variables, allant de la simple génération de texte, à des usages plus complexes (animation de réseaux de comptes *via* l'IAg, automatisant l'analyse de publications et la génération de réponses en retour, à base d'agents IA[3]). L'utilisation de l'IAg pour la génération de contenus textuels est de nature à faciliter, en théorie, le déploiement de manœuvres informationnelles sophistiquées, en permettant à la fois une plus grande diversité dans la reformulation d'un même texte ou d'un narratif, ainsi qu'en perturbant l'identification du caractère inauthentique et coordonné d'une campagne. Cela permet notamment d'éviter les écueils du procédé du *copy-pasta*[4], facilement détectable et pénalisé par les plateformes.

Enfin, ces technologies permettent de **simplifier la génération de contenus visuels, audio, ou vidéo**, avec un niveau de stylisation et de qualité susceptibles d'attirer l'attention et de générer de l'engagement. Bien que de nombreux contenus synthétiques diffusés sur les réseaux sociaux témoignent d'un usage récréatif de l'IA, en s'inscrivant dans le registre de l'humour ou de l'ironie[5], les observations de VIGINUM témoignent également d'un recours croissant à ces technologies dans le cadre de manœuvres informationnelles malveillantes (cf *infra* section I.3).

Ces gains de productivité en matière de génération de contenus permettent également aux acteurs de la menace informationnelle d'inscrire leurs stratégies sur le moyen et le long terme. A titre d'exemple, *Newsguard* a identifié 1 150 sites *web* d'information, entièrement ou principalement générés par des modèles de langage d'IA dans un total de 16 langues différentes[6]. Outre le fait que ces sites constituent des dispositifs pré-positionnés susceptibles d'être activés dans le cadre de manœuvres informationnelles, ils permettent également à leurs opérateurs de capter de grands volumes d'investissements publicitaires programmatiques[7] au détriment de médias authentiques, et ce par le biais de différents procédés (articles « pièges à clics », vidéos en lecture automatique, publicité contextuelle…).

b. Des capacités décuplées pour la réplication et la publication coordonnée de contenus inauthentiques à grande échelle

La facilité de création de contenus *via* les technologies d'IAg décuple ainsi l'aptitude des acteurs malveillants à mener des actions coordonnées à grande échelle sur plusieurs plateformes, voire même contribue à internationaliser leurs stratégies. En effet, outre la capacité de produire automatiquement différentes versions d'un même texte avec de légères reformulations pour tromper les mécanismes de détection, les technologies d'IA facilitent également la rédaction de contenus dans de nombreuses langues, dans une syntaxe convaincante, permettant de cibler plus facilement de nouveaux pays ou de

[2] Un prompt désigne une instruction ou un ensemble de données fournies à un système d'IA, lequel se sert de ces informations pour produire des réponses ou des créations sous forme de texte, d'images ou d'autres types de médias.
[3] https://huggingface.co/blog/ethics-soc-7
[4] Bloc de texte ou visuel copié-collé à l'identique ou presque, sur une ou plusieurs plateformes *web*, dans le but d'amplifier la visibilité d'un message.
[5] https://www.francetvinfo.fr/replay-radio/le-vrai-du-faux/macron-en-eboueur-trump-interpelle-obama-et-merkel-a-la-plage-les-images-creees-par-des-intelligences-artificielles-sont-de-plus-en-plus-realistes_5712170.html
[6] Newsguard AI Tracking Center, consulté le 7 janvier 2025, https://www.newsguardtech.com/special-reports/ai-tracking-center/
[7] Processus automatisé d'achat et de vente d'espaces publicitaires numériques.

viser plus spécifiquement certaines communautés linguistiques (diasporas par exemple) au sein d'un même pays.

c. Une aide pour la génération et la gestion de comptes inauthentiques sur les plateformes en ligne

L'un des défis techniques rencontrés par les acteurs de la manipulation de l'information relève de l'optimisation de la diffusion des contenus sur les plateformes en ligne. En effet, une opération d'ingérence numérique étrangère a besoin de vecteurs pour atteindre l'audience visée. Or, la gestion et l'animation des comptes inauthentiques, utilisés pour la diffusion, soulèvent à la fois des problématiques de crédibilité des profils, de furtivité et d'anonymisation de l'activité.

Comme l'a rappelé un rapport de la société *OpenAI*[8], l'utilisation d'images générées par une IA est massivement employée dans le cadre de campagnes de manipulation de l'information, afin de fournir à de faux comptes une photo de profil crédible, mais également pour industrialiser et crédibiliser la génération de comptes inauthentiques : nom, biographie, âge, contenu etc.

La combinaison de modèles génératifs textuels et de modèles génératifs visuels est ainsi susceptible de faciliter grandement la création de comptes inauthentiques, et ce malgré les dispositifs de détection des comportements inauthentiques mis en place par les plateformes elles-mêmes.

3. Cas d'usages observés de l'IAg à des fins d'ingérences numériques étrangères

S'appuyant sur les travaux de VIGINUM et d'acteurs de référence internationaux, cette partie vise à documenter différents cas d'usage observés de l'IAg à des fins d'ingérence numérique étrangère, sur la base d'investigations en sources ouvertes.

a. Principaux modes opératoires documentés

Dans le cadre des missions qui lui sont dévolues, VIGINUM a observé depuis 2022 un recours croissant à l'IA générative par les acteurs étrangers de la menace visant le débat public numérique francophone.

i. La génération de textes

Ce mode opératoire consiste à produire à grande échelle du contenu textuel *via* des applications d'IAg. À titre d'exemple, VIGINUM a observé à plusieurs reprises au cours de l'année 2024, notamment dans le contexte des Jeux Olympiques et Paralympiques de Paris, la diffusion de textes très probablement générés par IA[9] sur différentes plateformes telles que *X* et *Tumblr*, par des écosystèmes de comptes aux caractéristiques inauthentiques, très probablement affiliés au mode opératoire *Spamouflage*[10]. Bien que ces manœuvres aient été massives en termes de volume de messages publiés, elles n'ont finalement généré que très peu d'engagement, notamment en raison du fait que les comptes inauthentiques émetteurs ne disposaient que d'une faible audience, atténuant *de facto* le risque d'impact de ces publications coordonnées.

[8] https://openai.com/global-affairs/an-update-on-disrupting-deceptive-uses-of-ai/

[9] A l'instar des deux exemples de publications mentionnés *supra*, VIGINUM a identifié au total une centaine de messages diffusés avec un narratif quasiment identique lors de cette manœuvre.

[10] Le mode opératoire *Spamouflage*, documenté pour la première fois publiquement par *Graphika* en 2019, est un dispositif d'influence pro-RPC dont l'objectif est de diffuser des narratifs favorables aux intérêts du Parti communiste chinois (PCC) auprès d'une audience internationale. Il repose sur des réseaux de comptes aux caractéristiques inauthentiques chargés de mener des manœuvres informationnelles sur une multitude de plateformes. Cf. https://www.graphika.com/reports/spamouflage.

Exemples de publications d'un réseau de comptes inauthentiques ayant diffusé du contenu textuel généré par IA

De la même manière, en 2024, l'éditeur de sécurité *CyberCX*[11] a observé sur *X* un réseau de comptes aux caractéristiques inauthentiques dont l'activité principale consistait à amplifier des sujets politiques clivants en utilisant l'IA générative comme mode opératoire prédominant. Les publications de ce réseau, identifiées par la même formule « *As an AI language model…* » en début de message, ont été très peu visibles et n'ont généré là aussi que très peu d'engagement.

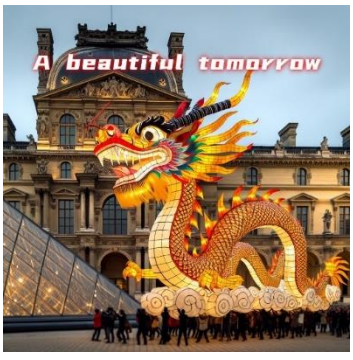
En outre, comme évoqué précédemment, la génération de textes par IA peut également être utilisée afin d'alimenter des sites de faux médias d'actualité en ligne grâce à la création de nombreux articles aux contenus manifestement inexacts ou trompeurs. A cet égard, plusieurs sites liés à l'écosystème *RRN*[12] sont alimentés par des articles générés par IA, lesquels sont ensuite diffusés auprès d'audiences françaises *via* de la publicité ciblée en ligne. L'ensemble des possibilités offertes en matière de génération de textes permet par ailleurs à ces acteurs de la menace d'optimiser leurs capacités de traduction de contenus, ce qui leur permet notamment *in fine* d'améliorer le référencement naturel de ces faux sites d'actualité.

ii. La génération d'images

VIGINUM observe également un recours croissant aux images générées par IA par les acteurs de la menace informationnelle. A la différence de la génération de textes, la finalité de l'usage d'images générées artificiellement semble être principalement d'illustrer un narratif donné par des images marquantes ou symboliques, plutôt que de véritablement tromper les utilisateurs sur leur authenticité.

Ces contenus, qui peuvent notamment imiter un style de dessin animé, utiliser des couleurs vives, ou bien utiliser des éléments visuels susceptibles de créer de l'émotion, cherchent à capter l'attention des internautes au sein de leur fil d'actualité, afin d'appuyer un narratif donné en vue de générer de l'engagement.

Le service a notamment observé la diffusion de visuels générés à l'aide de l'IA par des comptes aux caractéristiques inauthentiques affiliés au mode opératoire *Spamouflage*. Ces images correspondent pour la plupart à des dessins et à des caricatures, avec quelques représentations réalistes de sites connus.



Captures d'écran de visuels générés par IA diffusés sur *X* et *Tumblr* par des comptes *Spamouflage*.

[11] Cf. https://connect.cybercx.com.au/Intelligence-Update-CCX-IU-2024-004.
[12] https://www.sgdsn.gouv.fr/publications/maj-19062023-rrn-une-campagne-numerique-de-manipulation-de-linformation-complexe-et

Sur *X*, certains comptes de *trolling*[13] ultra-conservateurs diffusent ainsi régulièrement des images générées par IA, comme l'a relaté l'entreprise publique de télévision suédoise *SVT* dans une investigation récente[14] menée sur deux comptes, dont l'un comptabilisant plusieurs centaines de milliers d'abonnés. Dans le cas d'espèce, ce mode opératoire permet aux comptes de *trolling* d'appuyer des narratifs polarisants (anti-immigration par exemple) et de générer plusieurs millions de vues sur *X*, à moindre coût.

iii. La génération de vidéos et d'audios

Ce mode opératoire consiste à produire du contenu vidéo ou audio synthétique, manifestement inexact ou trompeur, afin de le diffuser sur plusieurs plateformes. Le contenu synthétique généré peut être original, ou bien proposer une version modifiée d'un contenu authentique, en altérant un ou plusieurs de ses éléments (voix, visage, apparence…).

Si les vidéos synthétiques semblent être de plus en plus utilisées sur les réseaux sociaux à des fins de divertissement d'audiences partisanes, le recours à des vidéos synthétiques crédibles dans le cadre de manœuvres informationnelles visant à tromper les utilisateurs semble pour le moment marginal. En effet, le coût induit par la fabrication d'un tel contenu (mixage de plusieurs technologies d'IAg, montage vidéo, incrustation de sous-titres, etc.) ainsi qu'une diffusion probablement moins aisée sur les réseaux sociaux, rendent ce mode opératoire moins accessible pour les acteurs malveillants.

Toutefois, le 13 février 2024, la cellule d'investigation de *France 24*, « Les Observateurs », a détecté et dénoncé la diffusion d'une vidéo manipulée par IA, usurpant l'identité de son et de son journaliste Julien FANCIULLI. Dans cette vidéo, le journaliste affirmait que les services de renseignement ukrainiens auraient eu pour projet d'assassiner Emmanuel MACRON et d'en faire porter la responsabilité à la Russie, de sorte à obtenir de nouvelles livraisons d'armes.

Capture d'écran du faux reportage de France 24 *sur* Telegram

Après analyse, « Les Observateurs » ont estimé que « le mouvement des lèvres du présentateur n'était pas synchronisé avec les propos prononcés dans la vidéo […] l'intonation de la voix semblait robotique » et que « certaines formulations […] ne correspondaient pas à la manière dont une information est donnée à l'antenne »[15]. Par ailleurs, le 13 décembre 2024, le média d'investigation indépendant russophone, *The Insider*, a publié un article présentant une nouvelle campagne du mode opératoire *Matriochka* (documenté en juin dernier par VIGINUM[16]) visant à convaincre les internautes que des professeurs issus d'universités prestigieuses appelaient l'Occident à lever les sanctions contre la Russie, tout en critiquant le président ukrainien Volodymyr ZELENSKY. Les investigations ont démontré l'utilisation d'outils d'IA dans ces vidéos, notamment pour cloner la voix des universitaires, dont certains d'entre eux avaient confirmé que les déclarations tenues n'étaient pas les leurs[17].

Capture d'écran d'une vidéo diffusée par le mode opératoire Matriochka

[13] Le *trolling* est un comportement en ligne où un individu cherche à créer des tensions, à provoquer ou à détourner une conversation de son objectif initial avec des messages offensants ou complètement hors sujet.
[14] https://www.svt.se/nyheter/utrikes/sa-gjorde-vi-granskningen-av-europe-invasion
[15] https://observers.france24.com/fr/%C3%A9missions/les-observateurs/20240214-une-tentative-d-assassinat-contre-emmanuel-macron-en-ukraine-attention-cette-vid%C3%A9o-est-truqu%C3%A9e
[16] https://www.sgdsn.gouv.fr/publications/matriochka-une-campagne-prorusse-ciblant-les-medias-et-la-communaute-des-fact-checkers
[17] https://theins.press/en/news/277174.

b. Regards croisés sur la menace informationnelle utilisant l'IAg : les acteurs internationaux institutionnels

Afin de dresser un panorama plus large de ces enjeux, cette partie intègre la perspective d'acteurs internationaux de référence en matière d'usage malveillant de l'IAg à des fins d'ingérence numérique étrangère.

i. Union européenne : point de vue du Service européen pour l'action extérieure

La division de la communication stratégique du SEAE dirige les travaux sur la désinformation étrangère, la manipulation de l'information et l'interférence. Elle a pour mandat d'analyser l'environnement informationnel, afin de permettre la mise en œuvre de la politique étrangère de l'Union européenne et de protéger ses valeurs et ses intérêts.

« La récente opération informationnelle ayant ciblé l'élection présidentielle et le référendum d'adhésion à l'UE en Moldavie a démontré que l'utilisation de l'intelligence artificielle (IA) pouvait être employée dans le cadre de manœuvres d'ingérences numériques étrangères. Dans ce cas précis, un contenu généré par l'IA et un logiciel dédié ont été utilisés pour tenter de tromper les électeurs, et d'influencer leur perception de l'adhésion à l'UE et de la candidate à la présidence Maia SANDU.

Une nouvelle forme d'engagement des utilisateurs : l'utilisation de *chatbot* basées sur l'IA

Le 29 septembre 2024, quelques semaines avant l'élection, un *chatbot Telegram* a été promu (@NuEuReferend_bot) pour mobiliser les citoyens moldaves afin qu'ils votent « NON » lors du référendum d'adhésion à l'UE. Selon les instructions de ce *chatbot*, après s'être inscrits à celui-ci, les utilisateurs se voyaient assigner des « tâches » à accomplir afin de convaincre d'autres citoyens de voter « NON », et recevoir en retour une rémunération financière. Le 29 septembre, ce *chatbot* a été promu sur *Telegram* par Ilian SHOR (@ilanshor), un politicien moldave sanctionné par les États-Unis et l'UE pour ses actions subversives à l'encontre de la démocratie en République de Moldavie (incluant notamment la fourniture de fonds illégaux pour soutenir l'activité politique locale pro-Kremlin, ainsi que « des connexions avec des oligarques corrompus et des entités basées à Moscou», selon un communiqué du département d'Etat américain datant de 2022), qui a annoncé la création du *chatbot* sur son compte *Telegram* le 29 septembre. D'après les observations du SEAE, c'est la première fois que, dans le contexte d'une élection européenne démocratique, une solution de *chatbot* qui offre un système de récompense financière a été utilisée pour distribuer du contenu invitant les utilisateurs à agir pour influencer le comportement des électeurs.

Détournement d'identité : usurpation d'identité à l'aide de technologies *deepfake*

De plus, au début du mois d'octobre, une vidéo générée par l'IA et contenant une imitation de la voix de la présidente de la Moldavie, Maia SANDU, a été diffusée sur *Telegram* et *TikTok*. Ce contenu *deepfake*[18] prétendait que l'adhésion de la Moldavie à l'UE obligerait le pays à adopter des lois concernant la communauté LGBTQ+, ou encore à vendre des terres nationales à des étrangers européens.

Le contenu a ensuite été amplifié par une chaîne *Telegram* affiliée à la version moldave d'un média russe, *Komsomolskaya Pravda*, et par le site moldova-news[.]com, qui appartient à Portal Kombat, un « réseau de propagande pro-russe structuré et coordonné », comme l'a rapporté VIGINUM en 2024[19].

Comme le démontrent ces incidents, le rôle de l'IA dans la manipulation de l'information soulève deux sujets de préoccupations majeures pour les acteurs impliqués dans la lutte contre la manipulation de l'information étrangère. D'une part, les solutions d'IA facilement accessibles peuvent faciliter la dissémination de contenus, et favoriser l'engagement des utilisateurs par le biais de messages sur mesure ciblant des individus ou des groupes spécifiques, rendant ainsi la manipulation non seulement répandue mais aussi personnalisée. D'autre part, les technologies d'IA telles que les *deepfakes* et la synthèse vocale réaliste ont le potentiel de brouiller les frontières entre le contenu inauthentique et le contenu réel, et d'éroder la confiance des citoyens dans les informations accessibles en ligne. »

[18] *Deepfake* : trucage audio ou vidéo à partir d'éléments existants, utilisant la technologie du *deep learning* pour changer le visage d'une personne dans une vidéo ou reproduire sa voix.
[19] https://www.sgdsn.gouv.fr/publications/portal-kombat-un-reseau-structure-et-coordonne-de-propagande-prorusse

ii. Suède : point de vue de l'Agence de Défense Psychologique

L'Agence de Défense Psychologique Suédoise (MPF) a pour mission de contrer les manipulations de l'information et les ingérences étrangères visant la Suède et ses intérêts.

« L'intelligence artificielle (IA), en particulier l'intelligence artificielle générative, transforme le paysage de la manipulation de l'information, permettant aux acteurs adverses de produire à très grande échelle une désinformation réaliste et peu coûteuse. Cela inclut notamment la génération de textes, d'images, de vidéos, de documents et d'extraits sonores fabriqués de toutes pièces, conçus pour manipuler les perceptions du public, éroder la confiance des citoyens et déstabiliser les institutions démocratiques. Des acteurs hostiles tels que la Russie, la Chine et l'Iran, ainsi que des groupes à motivation idéologique, exploitent l'IA de manière croissante pour créer des campagnes de désinformation sur mesure qui trouvent un écho auprès de publics spécifiques, en s'appuyant sur des modèles d'IA *open source* pour en contourner les restrictions.

L'un des principaux défis consiste à atténuer et à limiter la diffusion de la désinformation générée par l'IA, qui se propage rapidement et efficacement par le biais des médias sociaux et d'autres plateformes numériques. Ce contenu va des « *deepfakes* » sophistiqués aux « *cheap fakes* », des manipulations plus simples conçues pour une diffusion virale. L'accessibilité croissante de l'IA a brouillé les frontières entre le contenu authentique et le contenu manipulé, créant ce que les spécialistes appellent le « dividende du menteur », où même les contenus authentiques peuvent être mis en doute comme étant faux. Ces évolutions compliquent le *fact-checking*, accentuent la polarisation de la société et posent des défis importants à la résilience démocratique.

L'IA est principalement utilisée pour le blanchiment d'informations et l'amplification du contenu. Par exemple, des entités russes ont créé des sites web ressemblant à des sites d'actualité européens et américains légitimes. Ces sites, qui font partie d'opérations d'information telles que « Döppelgänger[20] », servent de « portails alimentés par l'IA » contrôlés par des intérêts russes. Les articles hébergés sur ces sites, dont le texte a été généré par l'IA, ont été disséminés *via* des réseaux de *bots* et des intermédiaires sur les plateformes de médias sociaux, et amplifiés par des réseaux coordonnés afin d'en maximiser la portée et l'impact. Ces opérations s'étendent souvent au-delà des médias sociaux, afin de rendre les campagnes de désinformation plus difficiles à identifier et à contrer.

Des acteurs étatiques et non étatiques motivés des convictions idéologiques ont exploité l'IA pour inciter à la division et amplifier les préjugés. Par exemple, des images générées par l'IA représentant des musulmans caricaturaux et menaçants ont été utilisées pour alimenter la haine envers les musulmans, et ces contenus ont été largement diffusés sur les médias sociaux par le biais de *hashtags*. Cette stratégie vise à normaliser les préjugés, et à les intégrer dans le discours dominant. De même, des clips vidéo manipulés, tels que ceux représentant des politiciens américains parlant chinois, peuvent manquer de sophistication mais restent des outils efficaces pour diffuser de la désinformation.

Parmi les autres exemples, peuvent être cités la génération de clips vidéo ou la production de contenus synthétiques visant à tromper les utilisateurs. Avant les élections américaines, les autorités ont identifié le groupe *Storm-1516* comme responsable de plusieurs campagnes d'influence très médiatisées, qui s'appuyaient sur de faux journalistes, de faux lanceurs d'alertes et sur des photos manipulées. En outre, les contenus synthétiques créés par des individus ont accentué la saturation de l'espace informationnel, rendant la détection et la lutte contre la désinformation de plus en plus difficiles.

Malgré ces avancées, la menace sous-jacente reste ancrée dans les tactiques d'influence traditionnelles adaptées aux nouvelles plateformes. L'IA a simplement amplifié ces techniques, les rendant plus généralisées, et plus difficiles à détecter. Pour faire face à ces défis, il est nécessaire d'avoir une population bien informée et résiliente, capable de résister aux manipulations adverses. La défense psychologique doit être un effort collaboratif impliquant les agences gouvernementales, les municipalités, la société civile, les organisations privées et le grand public. Il s'agit notamment de sensibiliser la population par le biais de campagnes ciblées, d'efforts d'éducation, de formation et la conduite d'exercices. La recherche démontre qu'une meilleure connaissance et une plus grande sensibilisation réduisent la vulnérabilité à la manipulation, renforçant ainsi les fondements d'une société ouverte et démocratique. »

[20] VIGINUM a documenté cette campagne sous le nom de *RRN*.

Alors que le paysage des menaces devient de plus en plus complexe, avec à la fois des acteurs étatiques et non étatiques qui utilisent l'IA pour influencer les conflits internationaux et les processus démocratiques, la défense psychologique doit évoluer. Pour renforcer ses capacités, le MPF a créé un *hub digital* pour analyser les risques, les vulnérabilités et les conséquences sur le paysage des plateformes numériques, y compris les réseaux sociaux, l'IA, les jeux vidéo, et les technologies émergentes. Cette initiative renforce la capacité du MPF à coordonner une stratégie de défense intersectorielle et cohérente à l'échelle nationale. L'exploitation des capacités défensives de l'IA permet d'apporter la réponse la plus efficace, garantissant la résilience des institutions démocratiques et la confiance de la société. »

iii. Canada : point de vue du Mécanisme de Réponse Rapide du Canada : cas d'étude Spamouflage

Rattaché au ministère canadien des affaires étrangères (Affaires mondiales Canada), le Mécanisme de réaction rapide (MRR) Canada est chargé d'échanger des informations et des analyses sur les menaces étrangères pesant sur les démocraties et d'identifier les possibilités de réponse coordonnée au sein du MRR du G7.

« Le Canada a observé l'évolution du paysage de la menace informationnelle au cours des dernières années. En particulier, le rythme rapide des progrès technologiques, et notamment l'intelligence artificielle (IA), a encore amplifié les menaces.

Les acteurs étatiques étrangers exploitent à la fois l'IA générative commerciale et interne pour la production de manœuvres d'ingérences numériques étrangères (FIMI), de manière dissimulée ou assumée. L'IA permet à ces acteurs de produire et de diffuser rapidement du contenu synthétique à grande échelle. Le but ultime est de semer la discorde, d'amplifier les désaccords et les griefs et de décrédibiliser les institutions gouvernementales.

Une tactique courante exploitée par les acteurs adverses étrangers est l'utilisation de l'IA pour produire des contenus ultraréalistes, notamment en créant souvent des avatars multilingues générés par l'IA pour créer et diffuser du contenu aligné dans plusieurs langues à grande échelle. En l'espèce, les expérimentations de micro ciblage ayant recours à des présentateurs générés par l'IA, utilisant différents accents et tonalités de couleur de peau pour attirer des audiences locales, sont très préoccupantes.

Une tendance tout aussi inquiétante est l'utilisation de l'intelligence artificielle par des États étrangers pour produire des fichiers audio et vidéo truqués, afin de diffuser des récits faux et trompeurs en ligne, dont certains peuvent être difficiles à détecter et à démystifier. L'utilisation de ces procédés dans le contexte d'actions violentes sexistes et identitaires, ou pour développer des contenus en ligne préjudiciables et abusifs destinés à cibler et à discréditer des individus, des militants, des journalistes ou des personnalités politiques considérés comme des menaces pour les acteurs étatiques, est particulièrement préoccupante.

Étude de cas : *Spamouflage*

En 2023, le Canada a enquêté sur une campagne probablement liée à *Spamouflage*, une opération déjà documentée par les acteurs impliqués dans la lutte contre la manipulation de l'information. Plusieurs personnes ont été ciblées par cette campagne, notamment des dissidents politiques, des parlementaires canadiens, le Premier ministre canadien et le chef du parti d'opposition canadien. La campagne a exploité les tactiques suivantes :

1. l'usurpation d'identité par le biais de vidéos générées par l'IA (*deepfakes*), afin d'accuser des dizaines de parlementaires canadiens de violations criminelles et éthiques ;
2. un réseau de *bots* tirant parti de la popularité des comptes vérifiés de médias sociaux de parlementaires canadiens ;
3. la publication par ce réseau de *bots* de milliers de commentaires en anglais et en français, amplifiant les accusations de fausses vidéos publiées dans le flux de la section commentaires sous les messages des comptes canadiens vérifiés.

Bien que l'impact de l'opération sur les parlementaires canadiens ait probablement été faible, le Canada reste préoccupé par l'utilisation de l'intelligence artificielle dans la création et l'amplification d'ingérences numériques étrangères visant à mener des actes de répression transnationale en ligne, et par le fait que ces activités deviennent plus persuasives et ont une plus grande portée. »

iv. Royaume-Uni : point de vue du bureau des Affaires étrangères et du Commonwealth

Rattachée au *Foreign, Commonwealth & Development Office* (FCDO) du Royaume-Uni, la direction des menaces cyber, informationnelles et technologiques a pour mission de fournir des informations et des analyses sur la menace informationnelle étrangère.

« En novembre 2023, les Etats participants au Sommet sur la sécurité de l'IA organisé par le Royaume-Uni à Bletchley Park ont reconnu le potentiel considérable de l'intelligence artificielle (IA) dans l'amélioration du bien-être, de la paix et de la prospérité dans le monde. Ils ont affirmé que l'IA devrait être utilisée de manière sûre, centrée sur l'humain, fiable et responsable pour réaliser pleinement ces avancées. Cet engagement souligne l'importance de la coopération internationale pour promouvoir une croissance économique inclusive, le développement durable et l'innovation tout en protégeant les droits de l'homme et en renforçant la confiance du public dans les systèmes d'IA.

La capacité de l'IA à manipuler l'information à grande échelle pourrait saper la confiance envers les institutions et les sources fiables, favorisant l'extrémisme politique et la polarisation de la société. Des acteurs malveillants, tant étatiques que non étatiques, exploitent l'IA pour manipuler les discours, polariser les débats nationaux et internationaux sur des sujets d'importance cruciale, et tenter de fragiliser les institutions démocratiques et la sécurité nationale. Des études publiques ont montré que les acteurs à l'origine des manipulations de l'information d'origine étrangère (FIMI) ont commencé à utiliser des outils d'IA générative pour créer du contenu textuel, audio, vidéo et des images.

L'IA offre également des opportunités pour contrer ces menaces en nous aidant à veiller et à comprendre les narratifs trompeurs dans plusieurs langues ainsi que les activités coordonnées des réseaux impliqués dans ces activités malveillantes. Cela permet également d'appuyer l'action gouvernementale en matière de lutte contre les manipulations et de soutenir l'intégrité de l'information et la stabilité nationale et mondiale.

En octobre 2024, le Royaume-Uni a sanctionné trois agences russes et trois individus clés impliqués dans des activités de déstabilisation visant l'Ukraine. Ces entreprises et leurs dirigeants sont responsables d'un vaste réseau en ligne, également connu sous le nom de *Doppelgänger*, qui inonde les réseaux sociaux de faux contenus, de documents contrefaits et de *deepfakes*, à destination des publics anglophone~~s~~, germanophone~~s~~ et francophone~~s~~. Le Royaume-Uni continuera de prendre des mesures contre les manipulations de l'information russes. »

II. Enjeux et perspectives de l'impact de l'IA sur la menace informationnelle

Si les acteurs de la menace informationnelle utilisent de manière croissante les technologies de l'IA générative pour armer leurs manœuvres, la question des effets et de l'impact de celles-ci sur l'opinion demeure néanmoins posée.

La mesure de l'impact d'une campagne numérique de manipulation de l'information ne fait actuellement pas l'objet d'un consensus au sein du monde académique. Principalement empirique, l'analyse de l'impact consiste souvent à relever des indicateurs quantitatifs de visibilité, fournis par les principales plateformes de réseaux sociaux (nombres de vues, de *likes*, de repartages ou de commentaires), mais ne fournissant qu'une vision parcellaire de l'exposition d'un lectorat ou d'un auditorat à la campagne, sans permettre la mesure de ses effets sur le long terme. Ainsi, et en dépit de quelques rares outils disponibles[21], l'impact des manipulations de l'information en ligne sur les comportements des publics visés ou exposés demeure mal connu.

Aussi, dans cette partie et sur la base de ses observations, VIGINUM s'attache à formuler des hypothèses de travail sur le risque d'impact des manœuvres malveillantes utilisant l'IAg.

1. Un risque réel d'élévation du niveau de menace informationnelle, mais un impact encore modéré à date

a. Un risque d'accroissement de la réactivité des acteurs étrangers malveillants

En réduisant le temps de production du contenu, les technologies d'IA sont de nature à accroître considérablement la réactivité des acteurs étrangers malveillants, en plus de leur permettre de produire à grande échelle. Sur la base d'une interaction avec un modèle de langage comme *ChatGPT*, il est possible de faire rédiger par l'outil du contenu de qualité mobilisable dans le cadre d'une opération d'ingérence numérique étrangère lancée sans préavis, dans une logique opportuniste par exemple. Cette automatisation de la création de contenus en temps réel *via* un modèle de langage ne nécessite aucune expertise particulière, facilitant ainsi la conception d'opérations simples à grande échelle.

b. Un risque d'accroissement de la furtivité des manœuvres informationnelles

Les modèles d'IA générateurs de contenus textuels sont aujourd'hui en capacité de produire des textes traduits en plusieurs langues étrangères, avec une qualité de syntaxe et d'usages parfois proches d'une langue maternelle. Au-delà de faciliter l'atteinte de certaines audiences (comme les diasporas par exemple), cette performance des outils d'IAg peut également contribuer à mieux dissimuler l'implication d'un acteur étranger dans le cadre de campagnes numériques de manipulation de l'information et rendre ainsi plus aléatoire la détection de ces dernières.

Par ailleurs, concernant la génération de texte, des tactiques aujourd'hui facilement identifiables – telles que le *copy-pasta* avec une traduction approximative - pourraient être remplacées demain par des procédés plus élaborés, *via* notamment la génération d'un grand volume de reformulations, plus complexes à mettre à jour. S'agissant des images, le recours à des générateurs d'image pour créer du contenu synthétique implique la mise à disposition de nouveaux types d'outils avancés de détection pour en identifier la source.

[21] Notamment : *Breakout Scale* de Ben Nimmo (The breakout scale: Measuring the impact of influence operations, 2020, https://www.brookings.edu/wp-content/uploads/2020/09/Nimmo_influence_operations_PDF.pdf) et *Impact risk-index* de EU Disinfolab (Towards an impact-risk index of disinformation: measuring the virality and engagement of single hoaxes, 2022, https://www.disinfo.eu/wp-content/uploads/2022/06/20220617_IndexImpactAssessment_Final.pdf)

c. Une évolution modérée de la menace à ce stade

Si l'IAg accroît la capacité des acteurs malveillants à produire de grands volumes de contenus sur les plateformes en ligne, cela ne semble pas constituer à ce stade une véritable rupture dans le champ des manipulations de l'information.

En effet, la qualité des contenus générés ne présume pas nécessairement de leur impact sur les audiences visées. Certaines recherches ont d'ailleurs démontré que des contenus manipulés exploitant des méthodes moins sophistiquées (« *cheapfakes* ») peuvent être aussi nuisibles que des contenus synthétiques sophistiqués[22]. À titre d'exemple, de vraies images, non générées par IA, mais instrumentalisées ou décontextualisées dans le cadre de manœuvres informationnelles, peuvent avoir un impact important[23].

En outre, si l'IAg permet d'accélérer les capacités de production et de dissémination de contenus et d'abaisser leur coût, elle ne permet pas encore de résoudre les défis liés à leur diffusion et leur viralité auprès de nouvelles audiences, qui demeurent le principal frein des manœuvres informationnelles[24] [25].

Aussi, et avec toute la prudence requise, l'intelligence artificielle semble aujourd'hui représenter davantage une « évolution » qu'une « révolution » en matière de menace informationnelle. Elle permet en effet une « industrialisation » des modes opératoires existants, avec des volumes de production bien supérieurs et à un moindre coût, mais ne semble pas encore avoir contribué à la création de modes opératoires inédits. Néanmoins, les évolutions technologiques rapides en matière d'IA présentent des risques structurels en termes de menace informationnelle à moyen et à long terme.

2. Perspectives d'évolution de la menace informationnelle liée à l'IA

a. Vers une altération du rapport à la réalité ?

La généralisation et l'accélération des capacités des modèles d'IAg font craindre le risque d'une prolifération massive de contenus générés par des outils d'IA sur les plateformes en ligne. Cette tendance peut déjà être observée sur certaines plateformes de réseaux sociaux (*YouTube*, *LinkedIn*, *TikTok*, *Pinterest*) et concerne également la production de médias synthétiques.

Sur le long terme, ce phénomène pourrait être susceptible d'accroître la méfiance des utilisateurs, et plus largement du grand public, à l'encontre de tout contenu en ligne, que celui-ci soit authentique ou non, et d'induire progressivement une forme de défiance envers la notion même d'information. L'essor de ces technologies pourrait ainsi provoquer un scepticisme généralisé (« *deep doubt* »), faisant peser pour les citoyens le risque d'une altération profonde de leur rapport à la réalité. C'est ce que constatent les auteurs d'un rapport de l'*Institute for Strategic Dialogue* (ISD) sur le rôle de l'IA dans les élections américaines de 2024 : « (…) *bien souvent, les contenus dont discutaient les gens n'étaient pas générés par l'IA, mais c'est le « spectre » de l'IA qui a eu un impact. L'IA donne l'occasion aux gens de nier la réalité comme ils le font déjà, mais de manière encore plus intense*[26] ».

---

[22] M. Hameleers, « Cheap Versus Deep Manipulation: The Effects of Cheapfakes Versus Deepfakes in a Political Setting », International Journal of Public Opinion Research, 2024, vol. 36, pages 1–9, https://doi.org/10.1093/ijpor/edae004

[23] A titre d'exemple, en octobre 2023, 250 pochoirs figurant des étoiles de David sont retrouvés sur des murs à Paris et proche banlieue. Des photographies de ces pochoirs ont ensuite été diffusées et amplifiées de manière inauthentique sur les réseaux sociaux. L'opération a été dénoncée par la France : https://www.diplomatie.gouv.fr/fr/dossiers-pays/russie/evenements/evenements-de-l-annee-2023/article/russie-nouvelle-ingerence-numerique-russe-contre-la-france-09-11-23.

[24] A. Narayanan et S. Kapoor, « *The LLaMA is out of the bag. Should we expect a tidal wave of disinformation?* », *AI Snake Oil*, 6 mars 2024, https://www.aisnakeoil.com/p/the-llama-is-out-of-the-bag-should

[25] A. Narayanan et S. Kapoor, « We Looked at 78 Election Deepfakes. Political Misinformation is not an AI Problem. », *AI Snake Oil*, 13 décembre 2024, https://www.aisnakeoil.com/p/we-looked-at-78-election-deepfakes

[26] « L'IA n'a pas chamboulé l'élection américaine, mais elle a changé le rapport à la réalité » de Radio-Canada. https://ici.radio-canada.ca/nouvelle/2120725/intelligence-artificielle-campagne-electorale-etats-unis-trump-harris

Ce même rapport de l'ISD[27] révèle par ailleurs que les utilisateurs s'appuient souvent sur des stratégies inadaptées pour déterminer si un contenu est généré ou non par l'IA. Au-delà de citoyens insuffisamment outillés, c'est toute la société qui fait face à l'enjeu critique de discerner la réalité de la fiction, et le synthétique de l'authentique.

Cette méfiance vis-à-vis de l'information pourrait être exploitée par des acteurs malveillants, comme l'ont théorisé des chercheurs de l'université de *Yale* aux États-Unis avec le syndrome de « *dividende du menteur* » [28] : plus une société apprend à être sceptique, plus il devient facile pour un menteur de remettre en question des faits pourtant irréfutables. Cela pourrait à terme déstabiliser en profondeur les processus démocratiques.

b. Risques liés à la prolifération de contenus synthétiques

Le recours à l'IA et la massification de la production de contenus permet également à certains acteurs malveillants de venir occuper des espaces en ligne jusqu'alors peu investis afin d'y diffuser leurs récits. En effet, la possibilité de produire des contenus à grande échelle autorise le déploiement de larges dispositifs, positionnés de manière durable et visant des audiences spécifiques. Il est ainsi probable de voir se multiplier des réseaux de faux médias locaux, à l'instar de l'opération PAPERWALL, dont le mode opératoire consiste à créer des sites imitant des portails d'informations locaux pour diffuser propagande et désinformation auprès de cibles localisées[29]. L'IA pourrait également être utilisée par des acteurs malveillants pour favoriser le référencement[30] de certains contenus sur des sujets très spécifiques.

Par ailleurs, sur des plateformes où l'algorithme de recommandation est particulièrement performant et vient proposer à chaque utilisateur des contenus similaires à ceux avec lesquels il a déjà interagi, tels que *YouTube* ou *TikTok*, la prolifération de contenus similaires sur une thématique donnée est susceptible d'accélérer le phénomène du « terrier de lapin » (« *rabbit hole* ») en enfermant les utilisateurs dans une bulle de contenus potentiellement manipulés.

c. Risque de pollution et d'auto-dégradation des modèles d'IA

La prolifération de l'utilisation de services d'IAg pose également la question des biais contenus dans les données d'entraînement utilisées, ainsi que celle des sources potentiellement mobilisées pour générer la réponse. S'agissant de l'entraînement des modèles, les fournisseurs de modèles pré-entraînés restent souvent opaques sur les données utilisées, ainsi que sur les potentiels biais dans les réponses des IAg. Les données d'entraînement sont issues de choix de collecte, de filtrage, et de traitement, qui se reflètent dans les générations des modèles. Une partie de la communauté scientifique travaille à une plus grande transparence de ces modèles[31], et mène des travaux afin de mesurer les biais dans les jeux de données utilisés[32], ainsi que dans les réponses des modèles d'IAg[33].

Pour éviter de fournir des réponses trompeuses, certains services s'appuyant sur des modèles d'IAg proposent des sources pour étayer leur réponse, mais sans analyse préalable de la qualité de l'information

---

[27] Rapport de l'*Institute for Strategic Dialogue (ISD)* intitulé « Déconnectés de la réalité : les électeurs américains face à l'IA et aux stratégies OSINT défaillantes », en date du 7 novembre 2024. https://www.isdglobal.org/digital_dispatches/disconnected-from-reality-american-voters-grapple-with-ai-and-flawed-osint-strategies/

[28] Kaylyn Jackson Schiff, Daniel Schiff, and Natalia S. Bueno, 2024, « *The Liar's Dividend: Can Politicians Claim Misinformation to Evade Accountability?* », *American Political Science Review*, https://doi.org/10.1017/S0003055423001454

[29] https://citizenlab.ca/2024/02/paperwall-chinese-websites-posing-as-local-news-outlets-with-pro-beijing-content/

[30] Le « référencement naturel » (en anglais *SEO* pour *Search Engine Optimization*), est une stratégie pour améliorer la visibilité d'un site web sur les moteurs de recherche afin d'atteindre un public cible.

[31] *The Foundation Model Transparency Index* de l'université Stanford https://crfm.stanford.edu/fmti/May-2024/index.html

[32] Par exemple, le projet *Knowing Machines* analyse les biais de construction du jeu de données LAION 5B, utilisé dans l'entraînement de certains modèles d'image : https://knowingmachines.org/models-all-the-way

[33] Par exemple, une analyse montre que les réponses apportées par le modèle d'intelligence artificielle Qwen 2 7B développé par *AliBaba* ne reflète pas la vision du monde occidental sur certains sujets politiques https://huggingface.co/blog/leonardlin/chinese-llm-censorship-analysis

de celles-ci. Ainsi, *Newsguard*[34] a démontré que certains services d'IAg utilisent comme source pour générer leurs réponses des sites du dispositif pro-russe *Portal Kombat* détecté et caractérisé par VIGINUM[35].

Outre la menace d'atteinte délibérée à l'intégrité des données, la multiplication des contenus générés en ligne sont susceptibles de polluer les données d'entraînement qui sont récupérées *via* de larges opérations de collecte du *web*. Selon certains chercheurs, la prolifération de ces contenus pourrait provoquer une forte dégradation de la performance des modèles d'IA, en étant intégrés de façon croissante à leurs données d'entraînement[36].

[34] https://www.newsguardtech.com/ai-monitor/french-language-ai-misinformation-monitor/
[35] VIGINUM, 2024, « Portal Kombat : un réseau structuré et coordonné de propagande prorusse », https://www.sgdsn.gouv.fr/publications/portal-kombat-un-reseau-structure-et-coordonne-de-propagande-prorusse
[36] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson et Yarin Gal, 2024, « *AI models collapse when trained on recursively generated data* », *Nature,* https://www.nature.com/articles/s41586-024-07566-y

III. L'IA comme opportunité en appui des opérations de lutte contre les manipulations de l'information

Si l'intelligence artificielle peut être utilisée de manière malveillante dans le cadre de campagnes de manipulation de l'information, elle peut aussi servir des causes plus nobles et notamment se montrer d'une redoutable efficacité dans l'analyse des comportements inauthentiques et autres modes opératoires.

A cet égard, les développements récents de l'intelligence artificielle, et notamment la mise à disposition de modèles pré-entraînés, offrent aux acteurs de la lutte contre les manipulations de l'information de nouvelles possibilités pour explorer, analyser et caractériser des profils sur des plateformes en lignes, des contenus textuels, audios, visuels ou vidéos, ou encore les dynamiques de diffusion d'un récit sur les réseaux sociaux.

1. L'expérience du service VIGINUM

a. Le Datalab de VIGINUM : l'analyse avancée de données au service des opérations

Afin de renforcer ses capacités d'investigation et d'analyse, VIGINUM a investi des ressources dans l'innovation technologique en se dotant, dès sa création en 2021, d'un Datalab. Au-delà d'une mission première d'appui aux investigations du service, le Datalab de VIGINUM mène une activité de recherche et développement (R&D), qui lui permet de mettre en place des méthodologies innovantes au service des missions opérationnelles, mais également de publier des articles académiques ainsi que, pour la première fois en 2025, du logiciel libre. L'activité de R&D du Datalab permet ainsi non seulement de fournir un appui technique et analytique aux investigations du service, et mais également d'outiller l'ensemble de la communauté mobilisée dans la lutte contre les manipulations de l'information (société civile, chercheurs, médias…).

Dans le cadre de ces travaux de R&D, le Datalab met en œuvre une large palette de méthodes et d'outils d'analyse de données, dont des solutions utilisant l'IA. Si l'IA permet en effet un gain significatif dans l'analyse (*cf infra*), celle-ci ne constitue néanmoins pas une solution miracle, et son utilisation nécessite une solide expertise technique, alliée à une bonne connaissance des forces et des limites de chaque modèle. Le Datalab adopte ainsi une approche pratique et pragmatique de l'IA, qui représente un ensemble de modèles et d'outils mobilisables en fonction des besoins, parmi un plus vaste ensemble de solutions.

De nombreux traitements de données mis en œuvre par le Datalab ne relèvent ainsi pas de l'IA. Par exemple, le coefficient de manipulation du trafic proposé par Ben NIMMO est un indicateur simple qui permet de mesurer l'amplification d'un *hashtag* sur la plateforme *X*, et qui ne repose pas sur de l'IA[37]. De la même manière, la méthodologie de détection de *trending topics* sur la plateforme *X* développée par VIGINUM repose sur de simples méthodes de détection d'anomalies statistiques[38].

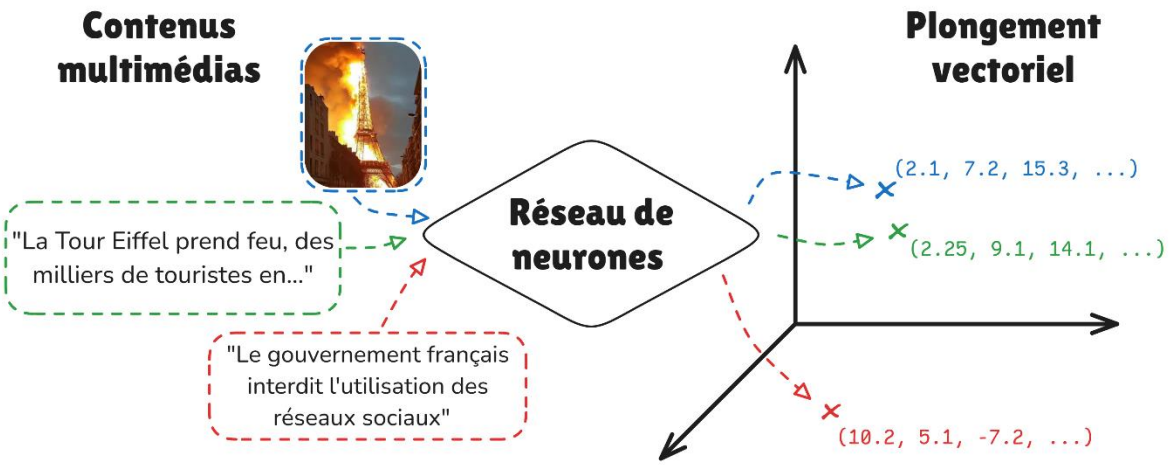
[37] Ben Nimmo, « Measuring Traffic Manipulation on Twitter. » Working Paper 2019.1. Oxford, UK: Project on Computational Propaganda. 35 pp., https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2019/01/Manipulating-Twitter-Traffic.pdf

[38] Cristian Brokate, Manon Richard, Lisa Giordani, and Jean Liénard. 2024. « SWATTING Spambots: Real-time Detection of Malicious Bots on X », Companion Proceedings of the ACM Web Conference 2024 (WWW '24)., Association for Computing Machinery, New York, NY, USA, 818–821. https://doi.org/10.1145/3589335.3651564

Exemples d'utilisation de l'IA mis en œuvre par le Datalab de VIGINUM :

***Utilisation directe d'un modèle via un* prompt**

Les modèles génératifs (*Mistral 7B instruct*, *Llama*, *Gemma*) étant pré-entraînés sur un corpus massif de textes, ils contiennent un ensemble de connaissances sur un grand nombre de domaines spécialisés. Cela leur confère de bonnes capacités en apprentissage *zero-shot*[39], qui leur permet d'être directement utilisés via des instructions (*prompt*). L'une des problématiques rencontrées avec ce type d'usages est la variabilité des réponses du modèle, ainsi que de leurs formats en fonction du *prompt* d'entrée[40]. Dans le cadre de ses opérations, le Datalab a recours à l'optimisation de *prompts* (*prompt engineering*), comme par exemple, pour détecter les pays mentionnés dans des publications issues de réseaux sociaux. Cette utilisation permet notamment de filtrer des publications ciblant spécifiquement la France au sein d'un corpus donné.



Un plongement vectoriel (ou embedding*) est une représentation de la sémantique des contenus dans un espace mathématique.*

Utilisation de plongements vectoriels (*embeddings*)

Un plongement vectoriel (ou *embedding*) est une représentation numérique d'un document (texte, image, audio, etc.) qui capture sa sémantique. Des modèles d'IA entraînés sur des corpus massifs de documents permettent ainsi de représenter des documents hétérogènes en une liste fixe de nombres dans un espace mathématique[41]. Cette représentation permet ensuite d'appliquer d'autres algorithmes, comme le calcul de distances entre documents, la détection d'images ou de textes proches sémantiquement ou la segmentation. Elle est utilisée par le Datalab dans sa méthodologie de détection des contenus dupliqués[42].

Ajustement (*fine-tuning*) de modèles

L'ajustement de modèles permet de spécialiser un modèle d'IA pré-entraîné sur une tâche et un jeu de données annotées. Cela permet d'utiliser les connaissances généralistes d'un modèle pour le spécialiser sur une tâche donnée. Le Datalab a par exemple utilisé ce principe pour classifier automatiquement les publicités dans le registre des contenus publicitaires de *Meta*[43].

[39] L'apprentissage *zero-shot* est la capacité d'un modèle d'IA à faire des prédictions sur des tâches ou des concepts sans jamais les avoir vus au court de son entraînement.

[40] Les réponses d'un modèle génératif peuvent parfois respecter la structure de réponse imposée par un *prompt* sans respecter la consigne.

[41] Par exemple, le modèle très connu BERT transforme une phrase en un vecteur de 768 nombres.

[42] Richard et al., « Unmasking information manipulation: A quantitative approach to detecting Copy-pasta, Rewording, and Translation on Social Media », 2023, https://arxiv.org/abs/2312.17338

[43] Voir les travaux présentés à la FOSDEM 2024 : https://archive.fosdem.org/2024/events/attachments/fosdem-2024-3204-detecting-propaganda-on-facebook-and-instagram-ads-using-meta-api/slides/22323/Fbads_FOSDEM_20240203103844_MlWExhL.pdf

b. Cas d'usage de l'IA pour la lutte contre les ingérences numériques étrangères

Le Datalab mobilise l'IA à la fois pour faciliter l'exploration des données, et pour identifier des marqueurs d'inauthenticité, afin de caractériser une ingérence numérique étrangère.

i. Analyse sémantique de contenus textuels

Dans le cadre de sa mission de détection et de caractérisation des ingérences numériques étrangères, VIGINUM analyse des données textuelles en langage naturel, telles que des messages collectés sur les plateformes de réseaux sociaux (*X*, *Facebook*, *Telegram*, *Threads*, etc*)* ou sur des sites *web* (blogs, faux médias, etc). Dans ce domaine, les progrès accomplis depuis dix ans en matière de traitement automatique du langage, et en particulier le développement de modèles de langage pré-entraînés, ont permis d'obtenir des gains significatifs dans de nombreuses tâches de traitement automatique du langage naturel, notamment concernant l'analyse sémantique de contenus textuels[44].

A titre d'exemple, dans le cadre d'un appui aux opérations, le Datalab utilise régulièrement des modèles de détection de thématiques (*topic modeling*), qui reposent sur des modèles de plongements vectoriels (*embeddings*), ainsi que sur des modèles de reconnaissance d'entités nommées[45].

Les modèles de *topic modeling* de type BERTopic permettent de détecter les sujets traités dans un corpus de textes, et de regrouper les contenus traitant du même sujet[46]. Ces méthodes permettent d'identifier les thématiques abordées par un acteur (« narratifs ») ou un ensemble d'acteurs dans un corpus, de déterminer si les sujets abordés sont d'intérêt pour le service, ou encore de cibler un sous-ensemble d'intérêt. Dans le cadre de l'appui aux investigations, la détection de thématiques constitue une brique importante de l'analyse exploratoire de contenus textuels. La veille technologique continue menée par le Datalab, notamment sur le sujet du *topic modeling*, permet au Datalab de se maintenir à l'état de l'art.

Au-delà de la détection de thématiques, le traitement automatique du langage permet également d'identifier des entités nommées au sein d'un corpus de contenus textuels. A titre d'exemple, le Datalab a mis en place une chaîne de traitement pour détecter les messages qui parlent d'un pays donné, que ce soit en le mentionnant en toute lettre, ou en mentionnant une ou plusieurs de ses localités. Cette chaîne de traitement permet notamment de repérer la publication de contenus concernant la France par des acteurs étrangers de la menace informationnelle. Parmi les méthodes testées, certaines s'appuient sur la réconciliation d'entités avec une base de connaissance existante, afin de relier les localités à un pays donné[47]. D'autres méthodes s'appuient sur des instructions (*prompts*) données à de grands modèles de langage (tels que *Mistral 7B*) pour identifier les pays associés aux localités mentionnées dans un texte.

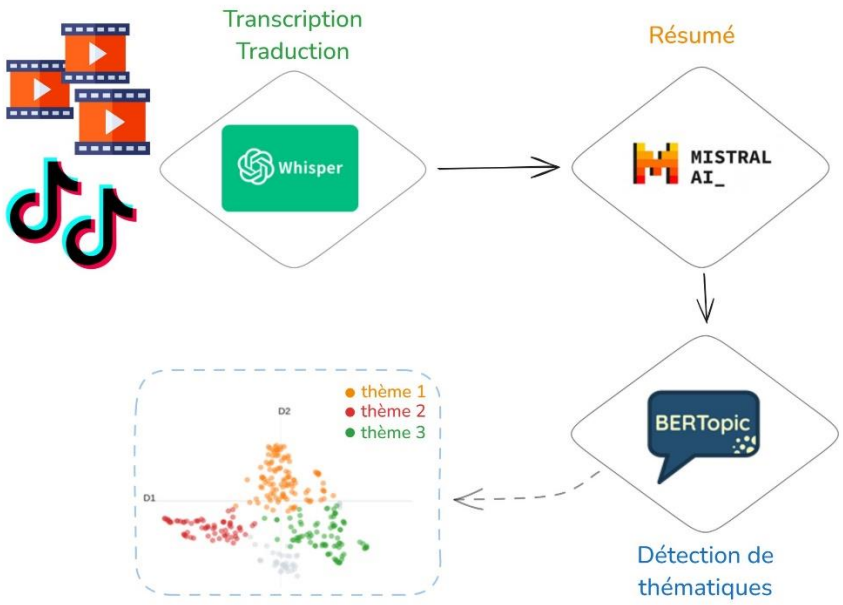
[44] Publié en 2013, le modèle Word2vec a rencontré un grand succès en permettant de projeter des mots dans un espace sémantique (Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, 2013, « Efficient Estimation of Word Representations in Vector Space, arXiv:1301.3781 [cs.CL]). A partir de 2018, les modèles de langage pré-entraînés comme BERT (Kenton, J. D. M. W. C., & Toutanova, L. K., 2019, « BERT: Pre-training of deep bidirectional transformers for language understanding », *Proceedings of naacL-HLT*, vol. 1, n°2.) puis GPT permettent une modélisation du sens des mots en fonction de leur contexte. Pour une histoire des modèles de langage jusqu'à ChatGPT, voir Pierre-Carl Langlais (7 février 2023), « ChatGPT : comment ça marche ? » *Sciences communes*, https://doi.org/10.58079/twxr

[45] La reconnaissance d'entités nommées est une tâche d'extraction d'informations qui cherche à localiser et classifier des éléments dans du texte en catégories prédéfinies comme un lieu, une personne, une organisation, une date, etc.

[46] https://github.com/MaartenGr/BERTopic

[47] Par exemple, la méthode Spacyfishing (https://spacy.io/universe/project/spacyfishing) permet d'associer une entité à la base de connaissance Wikidata.

ii. Construction d'une chaîne de traitement pour explorer le contenu de vidéos



Chaîne de traitement combinant différents modèles d'intelligence artificielle pour explorer le contenu d'un corpus de vidéos.

L'analyse exploratoire de contenus vidéos, issus par exemple des plateforme *TikTok* ou *YouTube*, est généralement plus complexe que l'analyse de contenu textuel pour les analystes. Il est néanmoins possible d'utiliser des grands modèles de vidéos et de langage[48] (*large language video models*) ou de grands modèles multimodaux (*large multimodal models*) pour analyser le contenu des vidéos, mais également d'utiliser des méthodes plus simples en se concentrant sur l'analyse de la bande-son de la vidéo.

En combinant différents modèles d'IA, le Datalab a par exemple mis en place une chaîne de traitement de données pour appuyer une investigation dans l'analyse exploratoire d'un ensemble de chaînes *YouTube* ou *TikTok*, et ce en utilisant uniquement l'analyse de la bande-son. Cette chaîne de traitement combine une brique de transcription, une brique de traduction, une brique de résumé automatique, et une brique de détection de thématique. Pour la transcription, il est possible d'utiliser des modèles *open source* comme la famille de modèles *WhisperAI*[49]. Cette brique peut ensuite être combinée à un grand modèle de langage pour obtenir un résumé automatique de la transcription de chaque vidéo.

Depuis 2023, de nombreux modèles de langage à poids ouverts, à l'instar de *LLama*[50], ou *open source* comme *Mistral7B*[51] ou *Mixtral8x7B*[52], peuvent être utilisés pour synthétiser le contenu des transcriptions. Les modèles de *topic modeling* (*cf supra*) peuvent alors être utilisés pour regrouper les vidéos traitant des mêmes thématiques. C'est la combinaison de ces modèles pré-entraînés qui facilite l'analyse exploratoire d'un corpus de vidéos.

[48] Par exemple https://llava-vl.github.io/blog/2024-04-30-llava-next-video/
[49] https://openai.com/index/whisper/
[50] https://github.com/meta-llama/llama3
[51] https://huggingface.co/mistralai/Mistral-7B-v0.1
[52] https://huggingface.co/mistralai/Mixtral-8x7B-v0.1

De l'importance *open source* et de la maîtrise de l'hébergement des modèles

Parmi les grands modèles pré-entraînés d'IA, certains sont fermés, et mis à disposition des utilisateurs uniquement sous la forme de service par l'entreprise qui les a développés. D'autres sont accessibles publiquement sous une licence *open source* à l'instar de Mistral 7B mis à disposition par l'entreprise *Mistral,* ou une licence plus restrictive comme les modèles *LLama* mis à disposition par *Meta* dont les poids sont téléchargeables mais les usages restreints par une licence spécifique[53].

Pour les services de lutte contre les manipulations de l'information, il est important de maîtriser les conditions d'hébergement de leurs outils pour des raisons de sécurité informatique, de sécurité opérationnelle, de protection des données à caractère personnel ainsi que pour des enjeux de souveraineté. VIGINUM s'est doté d'une infrastructure informatique en propre pour pouvoir héberger les données et réaliser les traitements dans un environnement informatique maîtrisé et sur des serveurs situés sur le territoire national.

Si les fournisseurs de services d'IA privés peuvent mettre en place des mesures de sécurité pour limiter les usages malveillants, il est plus difficile de sécuriser l'usage de modèles téléchargeables tels que les modèles *open source*. En revanche, le développement de modèles *open source* ou de modèles à poids ouverts[54] est un enjeu majeur pour permettre aux acteurs et entités mobilisés dans la lutte contre la manipulation de l'information d'utiliser des modèles pré-entraînés dans un environnement informatique maîtrisé.

Au-delà de l'analyse exploratoire de contenus, les outils d'IA sont également utilisés pour identifier des marqueurs d'inauthenticité, par exemple *via* l'identification de contenus massivement dupliqués.

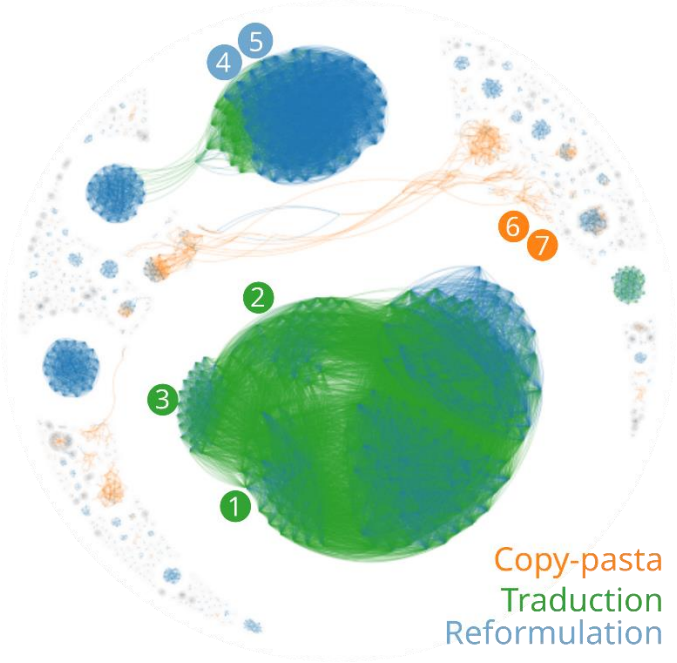
[53] Comme le souligne le PEReN, la définition de l'*open source* dans le domaine de l'IA n'est pas encore consensuelle. L'*Open Source Institute* (*OSI*) a proposé une définition en octobre 2024 mais peu de modèles sont conformes à la définition de l'*OSI* (https://www.peren.gouv.fr/actualites/2024-10-29_comparateur_iag_open_source/).

[54] L'*OSI* définit les modèles à poids ouverts (*open weights models)* des modèles dont les poids sont diffusés sous une licence libre sans que le code ou le jeu de données d'entraînement soit disponible (https://opensource.org/ai/open-weights) .

iii. Détection de duplications massives de contenus

Analyse d'un corpus de messages à l'aide de la méthode 3-delta.

VIGINUM a appliqué la méthode 3-delta au corpus « Venezuela 2021 » mis à disposition par le Twitter Transparency Center[55] et identifié des groupes de messages similaires.



Traductions :

- 1 : « ¿Cuándo cumplirá Cabo Verde con la orden de Ecowas [...] »
- 2 : « when will Cabo verde comply with Ecowas order [...] »
- 3 : « Cabo verde acatará a ordem do Ecowas [...] »

Reformulations :

- 4 : « #TWDxSTARChannel ya falta muy poco tiempo para el gran momento »
- 5 : « #TWDxSTARChannel ya por favor que el tiempo pase mas rapido solo por hoy #TWDxSTARChannel »

Copy-pasta :

- 6 : « Hoy toda Venezuela es Alex Saab exigimos liberen [...] de nuestro embajador Alex saab free Alex saab https://t[.]co/M1IJn8JEwg alex Saab »
- 7 : « Hoy toda Venezuela es Alex Saab exigimos liberen [...] de nuestro embajador Alex saab https://t[.]co/msEQgzrfFK alex Saab »

Pour diffuser un message textuel, les acteurs de la menace informationnelle peuvent avoir recours à différents modes opératoires : la duplication massive ou « *copy-pasta* » ; la reformulation ; ou encore la traduction. Pour détecter ces trois modes opératoires, le Datalab a développé la méthodologie 3-delta[56]. En utilisant un modèle de langage pré-entraîné, *Universal Sentence Encoder*[57], la méthodologie 3-delta détecte les paires de messages dont la proximité sémantique est forte. La méthode est ainsi en mesure

[55] En 2021, *Twitter* (aujourd'hui *X*) a identifié et supprimé un réseau de 277 comptes qui ont diffusé de manière massive des narratifs favorables au gouvernement vénézuélien. Le jeu de données avait été publié par le *Twitter Transparency Center*. Voir https://fsi.stanford.edu/news/twitter-takedown-december-2021

[56] Richard et al., « Unmasking information manipulation: A quantitative approach to detecting Copy-pasta, Rewording, and Translation on Social Media », 2023, https://arxiv.org/abs/2312.17338

[57] Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., ... & Kurzweil, R. (2019). « *Multilingual universal sentence encoder for semantic retrieval* ». arXiv preprint arXiv:1907.04307.

d'identifier les paires avec une forte proximité graphique et une forte proximité sémantique considérées comme du *copy-pasta*, les paires avec une forte proximité sémantique, une même langue et une moindre proximité graphique comme de la reformulation, et enfin, les paires avec une forte proximité sémantique, une moindre proximité graphique et une langue différente comme de la traduction. Cette méthode permet de détecter des groupes de messages qui sont anormalement similaires, et peuvent donc avoir été dupliqués de manière inauthentique.

Dans le cadre du Sommet pour l'action sur l'IA, organisé par la France les 10 et 11 février 2025, VIGINUM publie pour la première fois sous une licence libre la bibliothèque logicielle D3lta[58], afin de permettre aux différents acteurs de l'écosystème impliqués dans la lutte contre les manipulations de l'information de réutiliser et d'améliorer cette méthode.

iv. Détection de bots

L'utilisation non déclarée de comptes automatisés (ou *bots*[59]) est un marqueur d'inauthenticité. Au-delà d'heuristiques de détection relativement simples, pertinentes pour des *bots* à régularité de comportement (heures et fréquences des publications, publications similaires, …), l'IA peut être utilisée pour classifier d'autres types de comptes comme des *bots*. Le Datalab a ainsi construit son propre classifieur de *bots* sur *X* en s'appuyant sur l'analyse du réseau de *followers* et de *following* d'un compte donné[60].

v. Détection de contenus générés par IA

Face à l'utilisation croissante de l'IA générative par les acteurs de la menace informationnelle pour créer du contenu, les acteurs de la lutte contre les manipulations de l'information doivent se doter d'outils leur permettant de détecter les contenus synthétiques, soit pour être en mesure de distinguer les contenus synthétiques de contenus authentiques, soit pour caractériser le mode opératoire utilisé. A l'instar des outils d'IA générative, les détecteurs de contenus synthétiques sont eux-mêmes des outils utilisant des méthodes d'IA. Par exemple, la méthode *Binocular*, qui permet de détecter des contenus générés par des grands modèles de langage, s'appuie elle-même sur des grands modèles de langage pour mesurer la probabilité qu'un texte ait été écrit par un humain ou une IA[61].

De nombreux détecteurs de contenus synthétiques existent, avec des performances généralement variables en fonction des types de contenus analysés. Cela implique souvent pour les utilisateurs de tester de multiples détecteurs pour déterminer le caractère synthétique d'un contenu donné. Aussi, en vue de simplifier le processus de détection de textes et d'images synthétiques pour le grand public, VIGINUM et le PEReN ont joint leurs compétences pour inviter chercheurs et experts à participer au développement d'un commun numérique consistant en un méta-détecteur *open-source* de contenus artificiels, construit sur la base d'exemples réalistes observés lors de campagnes de manipulation de l'information[62].

[58] https://github.com/VIGINUM-FR/D3lta
[59] *Bot* : programme informatique automatisé pour simuler le comportement humain sur les réseaux sociaux. Un *bot* est capable de faire des publications, de laisser des commentaires, de partager ou d'aimer d'autres publications.
[60] L'approche s'appuie sur une méthode de réseaux de neurones convolutifs sur graphes hétérogènes développée par Feng et al., 2021, « BotRGCN: Twitter Bot Detection with Relational Graph Convolutional Networks », arXiv:2106.13092 [cs.SI]
[61] Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., ... & Goldstein, T. (2024). Spotting llms with binoculars: Zero-shot detection of machine-generated text. arXiv preprint arXiv:2401.12070.
[62] https://code.peren.gouv.fr/open-source/ai-action-summit

2. Les expériences issues de la société civile

Visant à compléter la perspective de VIGINUM, cette partie détaille des cas d'usage de l'IA mis en œuvre par des acteurs de la société civile impliqués dans la lutte contre les manipulations de l'information.

a. Point de vue de Lupa, agence brésilienne de fact-checking

« La *Lupa Newsroom*, une organisation brésilienne qui lutte contre la désinformation et mène des activités de *fact-checking*, fêtera ses dix ans en 2025. Au fil des ans, *Lupa* a été témoin de la sophistication croissante de la désinformation, tant dans ses stratégies que dans sa rapidité et son ampleur. Ces changements ont créé de nouveaux défis pour les *fact-checkers* et les ONG impliquées dans la lutte contre la manipulation de l'information, qui doivent désormais fournir des informations vérifiées au public plus rapidement que la désinformation ne se propage - car c'est pendant ce laps de temps que les « fake news » causent le plus de dégâts.

Mais il ne s'agit pas seulement d'une question de rapidité. Face à la sophistication croissante de la menace informationnelle, les outils doivent également être renforcés pour améliorer les capacités de *fact-checking*, notamment pour analyser, extraire des données, et produire des analyses de contenu approfondies dans un océan de plus en plus pollué par les informations trompeuses.

C'est dans cette optique que *Lupa* développe depuis 2023 son propre outil de veille du discours public. Grâce à *LupaScan*, les journalistes et les chercheurs peuvent analyser les déclarations de 596 élus fédéraux brésiliens (président, vice-président, sénateurs et députés fédéraux). En 2024, l'outil est devenu encore plus robuste, en intégrant plusieurs fonctions basées sur l'intelligence artificielle (IA). Les utilisateurs ont désormais accès à un tableau de bord qui présente automatiquement des graphiques et d'autres données visuelles, ce qui leur permet d'obtenir des informations pour leurs rapports et leurs recherches. L'une des nouvelles fonctionnalités comprend la reconnaissance faciale assistée par l'IA, et le regroupement des photos publiées par les personnalités politiques, ce qui permet d'analyser un vaste ensemble de publications comportant des images, même lorsque le nom de la personne représentée n'est pas mentionné dans le message.

L'une des fonctionnalités les plus remarquables est l'outil de transcription en temps réel. Cet outil a permis à *Lupa* de gagner en agilité et de réduire le risque d'erreurs lors de la couverture d'événements tels que les débats et les interviews, car la transcription manuelle n'est plus nécessaire. L'outil de transcription fonctionne en portugais et en anglais et sera bientôt étendu à d'autres langues.

Cet outil ne sera pas limité à la salle de rédaction de *Lupa*. L'organisation prévoit d'établir des partenariats, notamment avec des organisations latino-américaines, car plusieurs pays de la région organiseront des élections au cours de l'année 2025. »

b. Point de vue de Full Fact, organisation britannique indépendante de fact-checking

« Développée par l'organisation britannique indépendante de *fact-checking Full Fact*, *Full Fact AI* fourni aux organisations de *fact-checking* des outils pour lutter contre la désinformation à grande échelle. Déployés dans 45 organisations à travers 26 pays et disponibles en anglais, arabe et français, ces outils aident les *fact-checkers* à veiller le discours public, à identifier les fausses affirmations, et à lutter contre la désinformation répétée.

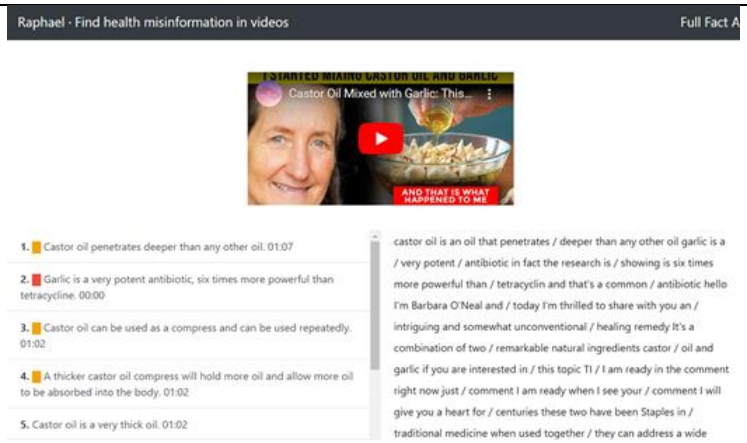
Les outils veillent diverses plateformes, notamment les sites d'information en ligne, les flux RSS, les archives parlementaires, les réseaux sociaux, la radio, la télévision, les podcasts et *YouTube*. Chaque jour, le contenu est converti en texte, découpé en phrases et enrichi d'informations supplémentaires. Ce processus aide les *fact-checkers* à gérer de vastes quantités de données à une échelle surpassant les capacités humaines.

Les données quotidiennes sont comparées aux vérifications précédentes, afin de détecter les informations trompeuses répétées, même si elles sont formulées différemment. Les *fact-checkers* peuvent alors rapidement adapter le travail existant et fournir des informations exactes.

L'une des solutions les plus récentes et les plus innovantes de *Full Fact AI* est son outil génératif de détection de la désinformation en matière de santé. Cet outil veille les contenus multimodaux sur toutes les plateformes en analysant les légendes, les images, les sons et le texte à l'écran. En classant les fausses informations sur la santé par ordre d'importance, l'outil permet aux organisations de hiérarchiser les ressources et de réagir rapidement aux allégations dangereuses, même lorsque les fausses informations sont implicites plutôt qu'explicites.

Les développements futurs comprennent l'automatisation de la veille des canaux enclins à la désinformation et la fourniture d'instantanés quotidiens, basés sur des mots clés, de la désinformation préjudiciable. Ces fonctionnalités amélioreront les flux de travail des *fact-checkers*, leur permettant de se concentrer sur la lutte contre les informations trompeuses les plus dangereuses.

Ces outils ne remplacent pas les *fact-checkers*. En combinant l'IA de pointe et l'expertise humaine, *Full Fact* peut accroître la portée et l'impact des *fact-checkers* sans compromettre l'exactitude. *Full Fact AI* démontre comment la technologie peut rationaliser les flux de travail, améliorer la précision et renforcer la confiance dans les sociétés. »



La capture d'écran montre le prototype Gen AI *de* Full Fact*, qui analyse les vidéos en ligne à la recherche d'informations trompeuses sur la santé, et fournit une liste de paraphrases d'affirmations pour permettre un examen rapide et facile par les* fact- checkers *de l'organisation*

Conclusion

Si l'IA offre des opportunités dans de nombreux domaines, elle représente un défi pour la lutte contre les manipulations de l'information, et fait peser un *continuum* de risques sur la sphère de l'information numérique.

D'une part, bien que les exemples documentés en source ouverte tendent à démontrer que l'usage de IA ne permet pas, pour le moment, de faciliter la propagation d'une campagne de manipulation de l'information ni d'en augmenter l'impact, le recours croissant à cette technologie pourrait entraîner une élévation structurelle de la menace informationnelle, en ce qu'elle permet d'accroître la réactivité des acteurs malveillants, ainsi que l'échelle et la furtivité de leurs actions.

D'autre part, par sa capacité à générer du contenu faux crédible, l'IA fait peser le risque d'un scepticisme généralisé du public à l'égard de l'authenticité de tout type de contenu en ligne. Si l'authenticité devenait plus facilement contestable, notre rapport à la réalité pourrait s'en trouver profondément altéré. Aussi, dans un contexte où la confiance du grand public envers l'information semble fragilisée, ce phénomène est susceptible d'accroître encore davantage la défiance de certains citoyens envers les médias et les *fact-checkers*, créant un terrain favorable aux acteurs de la manipulation de l'information.

Face à ces risques sans précédent, l'IA elle-même représente une solution efficace pour renforcer la lutte contre les manipulations de l'information, que ce soit pour explorer des données massives et variées, ou pour détecter des comportements inauthentiques. Afin d'augmenter les capacités de défense collectives, il est ainsi essentiel que tous les acteurs engagés dans cette lutte puissent innover, coopérer, et partager des outils d'analyse performants.

C'est le parti pris de VIGINUM, qui a décidé de publier pour la première fois – dans le cadre du Sommet pour l'action sur l'IA des 10 et 11 février 2025 – la bibliothèque logicielle D3lta sous une licence libre, afin de permettre aux différents acteurs d'utiliser cette méthode et de l'améliorer.

À PROPOS DE VIGINUM

Créé le 13 juillet 2021 et rattaché au SGDSN, le service de vigilance et de protection contre les ingérences numériques étrangères (VIGINUM) a pour raison d'être la protection du débat public numérique touchant aux intérêts fondamentaux de la Nation.

Ce service technique et opérationnel de l'État a pour mission de détecter et caractériser les campagnes de manipulation de l'information sur les plateformes numériques, impliquant des acteurs étrangers dans le but de nuire à la France et à ses intérêts.

Service de vigilance et protection contre les ingérences numériques étrangères | SGDSN

Crédit photo couverture : *photo and machines* sur Unsplash