INITIATIVE FOR MEDIA INNOVATION

DÉCRYPTAGE

2024: Perspectives d'une année d'élections mondiales à l'ère de l'IA

Gilles Labarthe

Edito

De Taïwan aux États-Unis, de l'Indonésie à la Finlande, du Sénégal au Venezuela, près d'une centaine de pays avaient rendez-vous avec les urnes en 2024, pour élire leur gouvernement. Mais plus que les enjeux démocratiques ou politiques, c'est la menace technologique qui faisait parler au moment d'entrer dans cette nouvelle année. IA, *deepfakes*, *bots* ou encore manipulation de données, les risques planant sur ces scrutins étaient nombreux, évoquant les scandales passés — de Cambridge Analytica au *deepfake* qui a fait basculer les législatives slovaques en 2023 — tout en y ajoutant des possibilités techniques nouvelles, toujours plus performantes.

Et si le danger pouvait paraître lointain vu de Suisse, c'était oublier les dérives observées lors des dernières élections fédérales, en 2023. Malgré l'appel commun de cinq partis nationaux à limiter l'usage de l'IA durant la campagne, on aura pu découvrir des affiches générées grâce à cette technologie, émanant de différentes formations politiques. Pire, en Argovie le Conseiller national UDC Andreas Glarner a généré un *deepfake* de sa rivale, Sibel Arslan. Une action pour laquelle le politicien sera au final condamné à une amende de… 3842,50 francs. Et de commenter : « *Das war mir der Spass wert* ». Littéralement, à ce prix-là, ça valait la peine de s'amuser.

En Suisse comme à l'étranger, il existe de nombreux exemples de manipulations d'opinions, d'utilisations abusives d'images modifiées ou de *deepfakes* bien plus travaillés. On pouvait légitimement craindre le pire pour 2024. Cependant, les premières études semblent s'accorder sur le fait que l'impact de l'IA fut limité, ne permettant pas par exemple à elle seule, de faire changer les résultats des scrutins.

Toutefois, l'influence de ces outils semblerait autrement plus insidieuse, déstabilisant de manière fondamentale la confiance du public dans l'information diffusée par les médias, quels qu'ils soient, des chaînes de télévision aux réseaux sociaux. Une fragilisation du rapport à l'information dont l'une des conséquences serait qu'une part croissante de la société s'enferme dans des convictions propres, véhiculées par des croyances ou des bulles informationnelles.

Face à ce péril, de nombreuses initiatives cherchent à mettre en place des alternatives et des outils, pour redonner au public les moyens de se retrouver dans cette masse d'informations ambiante. Qu'elles émanent du monde académique, des médias, d'associations, de fondations, voire des gouvernements eux-mêmes, ces initiatives ont joué un rôle durant cette super-année électorale.

Des élections présidentielles aux États-Unis, en passant par les législatives en Inde ou en France, décryptage de quelques-unes d'entre elles, avec celles et ceux qui les ont imaginées.

Christophe Schenk, président du conseil scientifique de l'IMI

Julie Schüpbach, IMI

INTRODUCTION

Perspectives d'une année d'élections mondiales à l'ère de l'IA

En 2024, un quart de la population mondiale s'est rendu aux urnes. Cette année électorale a révélé les risques de surmédiatisation de nouveaux outils mis au service de stratégies d'influence et de certains scrutins au détriment d'autres. Il est temps de reposer la question des moyens de détection des deepfakes et de la régulation de l'IA en politique, de façon globale et au-delà des pays occidentaux.

Ces derniers mois, de nombreuses études ont été publiées par des universités américaines autour de l'impact potentiel de l'IA générative sur les processus électoraux. Après des avertissements alarmistes proférés en 2023—2024 et au début de la course à l'élection présidentielle aux États-Unis, le constat général des scientifiques est devenu plus nuancé et prudent. Certes, les risques de diffusion de faux scandales politiques et de contenus trompeurs ne sont pas à négliger, en particulier lorsqu'ils sont publiés de façon ciblée à l'approche du jour du scrutin, ne laissant pas assez de temps pour les détecter et effectuer le travail de *fact checking* – aux États-Unis, on appelle cela le syndrome des « *October surprises* ». Mais si des *deepfakes* (surtout vidéos, images, son, mais aussi textes) parfois très convaincants ont bien été créés massivement avec l'IA générative, leur impact réel à des fins de persuasion ou manipulation des électeurs dans le cadre de campagnes politiques reste limité. Il serait moins important que ce que l'on aurait pu redouter. Il y a lieu cependant de différencier plus nettement la médiatisation de ce type de menace, sa perception par le public, les phénomènes renforçant le sentiment de « *AI panic* », et les effets observables et mesurables, dans le fonctionnement concret d'un système politique [1, 2]. Des effets secondaires ont aussi été signalés, par exemple le problème de frontière de moins en moins claire pour le public entre informations factuelles et vérifiées, et *deepfakes*. La distinction devient d'autant plus floue que des politiciens comme Donald Trump, réélu à la présidence le 5 novembre 2024, qualifient de plus en plus systématiquement de *deepfakes* de vraies informations qui les concernent, quand elles les présentent sous un jour défavorable [3].

Toutefois, l'inquiétude des experts porte davantage sur la capacité toujours plus grande de production, diffusion de masse et même de personnalisation, à moindres frais et à grande échelle, de contenus déstabilisants, et le manque de garde-fous. Nous serions confrontés davantage à une différence de degré de désinformation — avec des capacités augmentées de manière inédite grâce à l'IA générative — qu'à des changements profonds dans le caractère même de la désinformation, dont les recettes en termes de propagande politique sont déjà très anciennes.

Sur le plan des moyens de désinformation déployés, trois principaux types de menaces ont été étudiés et documentés, pointant des problèmes spécifiques. D'abord, le recours à l'IA et aux *deepfakes* pour dégrader l'environnement informationnel, augmentant les sentiments de polarisation et défiance envers les partis gouvernementaux, les autorités et les médias. Ces stratégies visent à amplifier les divisions, favoriser des formes de dé-cohésion sociale, et persuader au final des électeurs de s'abstenir de voter, ou d'agir à l'encontre de leurs intérêts et de leurs valeurs (voir aussi chapitre 2). Ensuite, des campagnes politiques menées *online* par microciblage alimenté par des données de profilage, l'IA et des *chatbots* véhiculant des messages personnalisés et à caractère émotionnel, sans présenter des informations de sources fiables. Ce type de microciblage devrait être davantage analysé et encadré, surtout lorsqu'il est utilisé à destination de jeunes électeurs. Enfin, se pose le problème de la concentration et centralisation de l'information par les géants du numérique (GAFAM) et par des entreprises technologiques d'IA générative — Meta, Google, OpenAI, etc. — avec entre autres les questions autour de l'absence de mention des sources, de modération du contenu, de repérage et dénonciation de partialité et de *hate speech*. Ce volet rejoint des préoccupations plus générales à propos du pouvoir que ces entreprises exercent sur les milieux politiques, ainsi que des limites de l'autorégulation et des réglementations internationales. Notons que ces recherches signalent également des utilisations positives des technologies d'IA générative, et des opportunités d'applications dans le domaine de la politique [4].

Nous avons tenté de synthétiser dans les pages qui suivent ces perspectives constructives. Et avons fait le choix d'ouvrir la discussion en incluant d'autres expertises et projets en cours, sur d'autres continents. Leurs apports respectifs vont dans le sens d'une meilleure complémentarité des approches proactives, réactives et collaboratives (voir chapitres 1 et 5), en matière de dispositifs contribuant à limiter les opérations de désinformation.

Beaucoup de professionnels rencontrés nous ont en effet rappelé l'urgence de promouvoir une réflexion globale réunissant toutes les parties prenantes. Il resterait à étendre ce champ d'études et d'interventions à une plus grande diversité de manipulations recourant à l'IA (y compris les *cheapfakes*), en constante évolution, exploitant les vides juridiques et les « zones grises » des réglementations existantes ; et aussi, à des contextes socioculturels très exposés mais encore trop peu étudiés : notamment les communautés défavorisées, minoritaires et issues de la diaspora (chapitre 2), et au-delà des pays occidentaux, les pays d'Asie (chapitre 4) et d'Afrique (chapitre 5).

Toutes ces contributions apportent des suggestions de « *best practices* » à l'intention des autorités politiques, des entreprises, des médias et de la société civile. La liste serait longue à établir. Nous en avons mis en évidence certaines des principales en fin de rapport (chapitre 6), dans la perspective d'ouvrir le débat, et des horizons pour de futurs projets.

CHAPITRE 1

Manipulations, deepfakes et contenus synthétiques ou la bataille des images

L'histoire des manipulations d'images à des fins de propagande est déjà ancienne, mais l'IA en a développé les possibilités sans commune mesure et popularisé l'accès, jusqu'à diffuser des contenus 100% synthétiques. Selon Touradj Ebrahimi, professeur à l'EPFL, nous sommes entrés dans une ère de nouveaux enjeux technologiques, légaux, sociétaux, commerciaux et éthiques. Les initiatives visant une meilleure régulation et le développement d'outils de détection des *deepfakes* restent toutefois limitées.

« LE MAILLON FAIBLE, C'EST L'ÊTRE HUMAIN »

Les images et informations visuelles ont un impact considérable sur nos émotions, nos sentiments. Elles sont aussi propices aux manipulations, et notamment au mélange – plus ou moins subtil – avec des messages cachés, subliminaux ou de désinformation. Le fait qu'elles représentent un des vecteurs préférés d'opérations de propagande constitue une raison supplémentaire de s'y intéresser de près. Au sein du laboratoire de traitement de signaux multimédia qu'il dirige à l'EPFL, Touradj Ebrahimi a intégré dans son champ de recherches les problématiques de confiance et sécurité dans l'imagerie basée sur l'IA, les innovations et transferts de technologies, ainsi que les initiatives concrètes visant à une meilleure régulation.

Les outils numériques au service de l'image peuvent susciter une certaine fascination, ne serait-ce que par des performances et capacités toujours plus élevées. L'expert relève néanmoins plusieurs sources d'inquiétudes. Par exemple, la profusion exponentielle d'images manipulées, le nombre de nouveaux outils à disposition offrant la possibilité de déléguer les tâches, les compétences nécessaires et le savoir-faire, et la variété d'acteurs diffusant toujours plus de contenus à caractère politique, via des canaux toujours plus nombreux. Touradj Ebrahimi attire aussi l'attention sur une autre réalité, confirmée par de nombreuses études (voir encadré) : « Dans la plupart des problèmes de sécurité, de *fake news*, le maillon faible, c'est l'être humain. En comparaison, il y a nettement plus de distribution de l'information erronée par des êtres humains, par des personnes mal intentionnées ou trop crédules, qui ne remettent pas les contenus en question… que par des *bots* ». La technologie seule n'explique pas l'accélération de production et de diffusion massive des *deepfakes*, entre autres sur des plateformes comme X (ex-Twitter, voir chapitre 3). Il y a bien des responsabilités et intentions humaines derrière. « Or, côté prévention, nous restons sous-équipés face à la gigantesque "bataille des images" qui se déroule en temps réel en période d'élections ou de propagande de guerre. Une chose est certaine, il existe aujourd'hui beaucoup plus d'outils utilisant l'IA pour créer des contenus synthétisés, que pour les contrer, les détecter, ou amener les informations nécessaires aux utilisateurs leur permettant de savoir qu'ils ont été créés par de l'IA. » Pourquoi ce rapport de force reste-t-il inégal ? « La raison en est très simple : commerciale. Les gens sont davantage prêts à payer pour des outils qui leur apportent du divertissement, que pour s'assurer une information correcte et de qualité. Nous ne pensons pas encore que la situation soit suffisamment inquiétante et critique », souligne l'expert. Comme il le constate : « Cela impressionne beaucoup plus de montrer ce que l'IA peut faire, en matière de création de contenus ludiques, étonnants, spectaculaires ou choquants, que des détecteurs. » Les médias ont aussi leur part de responsabilité, relève au passage Touradj Ebrahimi : fallait-il autant attirer l'attention du public sur les versants sensationnalistes des *deepfakes* — comme lors du duel opposant Donald Trump à Kamala Harris — plutôt que le sensibiliser sur le besoin de vérification des sources et de décryptage ?

TROIS APPROCHES

En matière de prévention, la perspective d'utiliser « l'IA contre l'IA » représente une piste prometteuse. « La quasi-totalité des détecteurs de *deepfakes* aujourd'hui, qui donnent de bons résultats, sont aussi basés sur l'IA, avec un auto-apprentissage. Beaucoup de ces outils mettent en compétition deux IA, l'une qui génère des contenus synthétisés, l'autre qui essaie de les détecter. On appelle cela le GAN, *Generative Adversarial Network* » (réseaux antagonistes génératifs). Touradj Ebrahimi regrette cependant le peu d'opportunités de développement de ce type de démarche sur le marché. Et un manque de clarté sur les approches utilisées, qui gagneraient à être mieux combinées.

\ **PROACTIVE**. Certains outils servant à détecter les *deepfakes* et contenus synthétisés par l'IA se basent ainsi « sur une approche proactive ou de provenance » (*provenance and traceability signalling*). Le professeur cite le cas de la Content Authenticity Initiative, lancée par Adobe [1]. « Le but de cette traçabilité est de permettre de savoir d'où vient un contenu, s'il a été manipulé et si oui, ce qui a été fait, de manière fiable, afin de le labelliser en tant que généré par l'IA. Le tout peut être vérifié par une signature numérique. Mais cette approche nécessite un large consensus au niveau international », avertit le spécialiste.

\ **RÉACTIVE**. Vient ensuite l'approche réactive, avec par exemple l'outil FakeCatcher, d'Intel [2] : face à un manque d'informations de provenance, on s'intéresse ici à la recherche d'anomalies, aux traces de ce que l'IA ne parvient pas à créer correctement. McAfee, entreprise de détecteurs de virus informatiques, a elle aussi commencé à détecter des *deepfakes*, mais ne propose encore qu'un seul outil qui fonctionne pour l'audio des vidéos analysées [3, 4].

\ **COLLABORATIVE**. Enfin, le *fact checking* représente une approche collaborative. Le but est d'effectuer des recherches à partir de bases de données sur les fausses informations, de croisements de sources ou d'archives. Que ce soit aux États-Unis, en France ou en Suisse, de nombreuses initiatives de *fact checking* collaboratif réunissant journalistes et citoyens ont vu le jour ces dernières années (voir chapitres 2 et 5).

MANQUE DE MOYENS

Dans le domaine de la sécurité, « aucune de ces trois approches n'est totalement sûre », prévient le chercheur de l'EPFL. « Il faut les cumuler, en mettre plusieurs couches, avec plusieurs outils pour diminuer les risques. Envisager non pas une solution unique, mais une combinaison de plusieurs outils utilisés, et une règle de majorité ». C'est la démarche suivie pour « la norme JPEG Trust, basée non pas sur une solution unique, mais sur une combinaison de plusieurs outils, avec plusieurs indicateurs de confiance. Il s'agit là d'une avancée importante, parce que beaucoup de *deepfakes* sont avant tout visuels ».

Les enjeux sont multiples, d'autant que dans la course aux innovations numériques, « c'est le jeu du chat et de la souris, ces détecteurs ont toujours un temps de retard par rapport aux nouveaux outils de création ». Un des défis sur lesquels travaille aujourd'hui son laboratoire : la détection de *deepfakes* récents, basés non plus sur des synthèses de contenus existants, mais créés de toutes pièces par l'IA générative, à partir de prompts : « Nous essayons de maximiser la probabilité de détection en tenant compte du *workflow*, c'est-à-dire, depuis la capture d'une image, de tous les outils de traitement utilisés, toutes les traces de codage, jusqu'à leur visualisation. Cela donne généralement de bons résultats ». L'expert déplore toutefois le manque d'investissements dans le domaine de la détection. Finalement, reste à questionner si les instances internationales ont pris la mesure de la situation.

LE SAVIEZ-VOUS ?

Touradj Ebrahimi cite les résultats surprenants de deux études récentes portant sur les *fakes news* et les *deepfakes*. La première a été publiée en 2018 par des chercheurs du Massachusetts Institute of Technology (MIT). En analysant 126 000 posts publiés sur Twitter entre 2006 et 2017 au sein d'un groupe échantillon de 1500 personnes, il a été possible de montrer que les fausses informations circulaient six fois plus vite que les vraies. Les *fake news*, à caractère politique notamment, sont aussi beaucoup plus partagées et re-tweetées, pouvant ainsi toucher jusqu'à cent fois plus de personnes. Parmi les conclusions de cette étude : ce sont bien les humains, et non les robots, qui sont les premiers responsables de la diffusion et rediffusion de contenus de désinformation [5].

Dans une autre étude menée par une équipe de l'Université de Californie (Berkeley) et de Lancaster, publiée en 2022 [6], des chercheurs ont soumis aux participants de vraies photos de visages humains, et des faux portraits synthétiques générés par StyleGAN2. Ils ont non seulement démontré que la distinction était devenue presque impossible à établir à l'œil nu entre les deux catégories, mais aussi, que « les *deepfakes* produits par l'IA et de haute qualité suscitent plus de confiance que de vraies images de personnes et personnalités », résume l'expert.

CHAPITRE 2

États-Unis Désinformation et enjeux technologiques

Au-delà des deepfakes, apparaît le problème plus vaste de la désinformation, parfois amplifiée par des outils d'IA, les filtres et algorithmes, témoigne Sisi Wei, l'une des co-fondatrices de The Markup. Cette spécialiste du datajournalisme d'investigation se focalise sur les impacts sociaux de certains usages des technologies, devenus préoccupants.

Sur un ton polémique, violent et démagogue, sans souci de véracité des affirmations avancées, Trump et ses partisans se sont distingués par des usages de l'IA paradoxalement beaucoup plus sommaires et amateurs que des *deepfakes* misant sur la qualité d'images de synthèse – *deepfakes* qui avaient pourtant été désignés par des experts comme principale menace en période d'élections. Dans la perception du public, il est devenu de plus en plus complexe durant la course à la présidence américaine de distinguer la limite entre images manipulées, fausses accusations de trucages, et faits objectifs et vérifiables. De nombreux professionnels des médias s'inquiètent ainsi « de la désinformation en général, n'ayant pas forcément besoin de s'appuyer sur de l'IA », relève Sisi Wei, journaliste et analyste, basée à New York.

LES APPROCHES COLLABORATIVES DE *FACT CHECKING*, D'INVESTIGATION PARTICIPATIVE ET L'ÉDUCATION AUX MÉDIAS FONT PARTIE DES SOLUTIONS PROPOSÉES

NOUVEAU REGARD

Comment lutter contre ces stratégies de désinformation, qui prennent souvent des tournures plus insidieuses ? Pour cette experte, les approches collaboratives de *fact checking*, le journalisme d'investigation et de données participatif, l'éducation aux médias et aux réseaux sociaux font clairement partie des pistes de solutions à proposer. Sisi Wei rappelle combien le recours à des compétences croisées lui a permis ces dernières années de développer « un autre regard sur notre manière d'appréhender le journalisme », ainsi que des applications numériques très concrètes au service d'une information d'intérêt public, comme pour Propublica, un média d'investigation américain ayant le statut d'ONG, souvent cité comme modèle du genre [1]. Sisi Wei a depuis contribué à lancer plusieurs « médias citoyens », comme The Markup [2] en 2019, ou la rédaction de CalMatters [3].

Suivant une démarche originale, les deux rédactions à but non lucratif de The Markup et CalMatters se donnent pour objectif de montrer comment les usages de la technologie par des multinationales, mais aussi par les grandes institutions, peuvent bel et bien avoir un impact sur nos sociétés, et sur des communautés minoritaires en particulier. La question des responsabilités (*accountability*) se retrouve au centre de leurs enquêtes. Leurs rédactions ont certes mené des ateliers de sensibilisation et de formation pour l'identification de *deepfakes* : par exemple, pour sensibiliser la diaspora vietnamienne en Californie [4] ; ou dans un contexte familial, pour prévenir les clivages entre générations [5] ; ou encore, pour contribuer à l'éducation aux réseaux sociaux, autour des problèmes liés aux biais et filtres sur ces plateformes. De fait, beaucoup d'investigations réalisées par The Markup se focalisent sur des menaces très concrètes. Il s'agit avant tout d'expliquer au travers d'exemples précis comment opère la désinformation, et les risques liés à la suppression d'informations importantes [6].

REVENUS DE LA DÉSINFORMATION

Certaines investigations récentes de The Markup montrent ainsi comment la désinformation se retrouve amplifiée par les réseaux sociaux. « Nous avons réalisé une enquête démontrant comment la violence politique profite à la société Meta et à Facebook, qui en retire des millions de dollars de revenus, au travers de publicités directement liées ». Pour ce faire, The Markup et CalMatters ont utilisé les modèles publicitaires de Facebook (combien coûte la publication et diffusion d'une annonce) pour calculer les montants des revenus concernés ; et quels types de publicités sont liées (pour la vente de t-shirt, d'une voiture, etc.) aux contenus évoquant par exemple la tentative d'assassinat de Trump en juillet 2024 [7] (voir aussi l'encadré).

Très axé sur le recueil, le traitement et la visualisation de données, le modèle de fonctionnement collaboratif de The Markup et de CalMatters pourrait se renforcer aux États-Unis et inspirer d'autres projets similaires en Europe et en Suisse (voir chapitre 5). Sisi Wei anticipe d'autres risques pouvant toucher les processus électoraux aux États-Unis, en 2025 : « En plus de la désinformation, je crains une évolution de la production et diffusion des *deepfakes* via des zones grises, notamment des émetteurs "lambda", "Monsieur et Madame tout le monde", dont on ne sait rien, et qui peut-être même n'existent pas, relayant des accusations à impact politique, obscures et invérifiables, mais produites et diffusées en sous-main par des réseaux organisés. La détection et le *fact checking* seraient encore plus difficiles à effectuer, alors que leur impact sur des communautés et l'opinion peut s'avérer très concret. Surtout dans le cadre des prochaines élections locales de 2025 ».

LE SAVIEZ-VOUS ?

Sur Facebook, de curieux produits de merchandising sont apparus juste après la tentative d'assassinat de Trump en juillet 2024. Son visage, avec ou sans oreille ensanglantée, est devenu omniprésent : sur des tasses à café, des chemises hawaïennes, des cartes à collectionner, des pièces commémoratives, ou des pendentifs en forme de cœur. Certains de ces portraits d'un Trump montré à son avantage, avec un visage combatif, mâchoire serrée et poing levé après les tirs, se basaient sur des images prises sur place par les photographes Doug Mills (pour le New York Times) et Evan Vucci (pour Associated Press). Mais l'équipe de campagne de Trump a aussi diffusé elle-même d'autres éléments et objets « commémorant sa survie ». La violence politique de cet événement a profité aux annonceurs en ligne : ils ont saisi là « une opportunité commerciale, diffusant des publicités sur Facebook à des partisans avides de produits dérivés », résume The Markup. « Dans les dix semaines qui ont suivi la fusillade, les annonceurs ont payé à Meta entre 593 000 et 813 000 dollars pour des publicités politiques mentionnant explicitement la tentative d'assassinat. Même si les dépenses consacrées aux publicités d'assassinats représentent une fraction des plus de 100 milliards de dollars de revenus publicitaires de Meta, la multinationale améliore aussi ses résultats financiers lorsque se produisent des tragédies, comme des fusillades de masse aux États-Unis, ou la guerre, au-delà. Après l'attaque du 7 octobre contre Israël l'année dernière et la riposte contre Gaza, Meta a connu une augmentation importante des dépenses liées au conflit, selon notre étude. Facebook a par ailleurs reconnu que le contenu polarisant et la désinformation sur sa plateforme ont déjà incité à des actes de violence réelle. »

CHAPITRE 3

France Législatives 2024 à l'épreuve de X et des réseaux d'ingérences

Un groupe de chercheurs du CNRS s'est intéressé à l'impact de la plateforme X et des tweets politiques français durant la période des législatives de l'été 2024. Leur étude montre que les réglages algorithmiques de la plateforme ont été utilisés afin d'amplifier la diffusion d'information contenant des messages toxiques, nuisibles au débat public, qui se retrouvent sur-recommandés aux utilisateurs.

David Chavalarias est directeur de recherche au CNRS et au Centre d'analyse et de mathématique sociales de l'École des hautes études en sciences sociales (EHESS). Il est également directeur de l'Institut des systèmes complexes de Paris Île-de-France. Son champ de recherches concerne « l'impact du numérique sur les sociétés, les démocraties, l'influence des réseaux sociaux sur la formation des opinions, les techniques de manipulations et stratégies en ligne », comme il le rappelle. Ces phénomènes sont devenus de plus en plus perceptibles et préoccupants depuis une dizaine d'années.

LA DIFFUSION D'INFORMATION PAR L'ALGORITHME DE LA PLATEFORME EST BIAISÉE SUR DES DIMENSIONS NUISIBLES AU DÉBAT PUBLIC, NOTAMMENT CELLE DE LA TOXICITÉ

« Avec le Brexit, notamment (référendum sur l'appartenance du Royaume-Uni à l'Union européenne, juin 2016, ndlr), nous avons vu pour la première fois une votation majeure être perturbée par des opérations en ligne. Je me suis dit que c'était alors le moment de lancer un observatoire au niveau français de la politique sur Twitter, qui était à l'époque une fenêtre intéressante puisque toutes les campagnes politiques françaises se faisaient en priorité sur ce réseau. Nous avons lancé avec un collègue, Maziyar Panahi, le Politoscope en 2016, avec une collecte qui recueillait tous les jours entre 500 000 et deux millions de tweets comprenant des messages politiques ».

PREMIER OBSERVATOIRE

Le Politoscope se base sur une méthodologie comprenant la modélisation, des études empiriques de terrain, un vaste travail de recueil de données, puis leur cartographie pour la visualisation de flux d'échanges (voir encadré). Premier observatoire de la politique sur Twitter au niveau français, il a vite servi pour s'intéresser à l'élection présidentielle de 2017. Parmi ses dernières recherches, l'équipe du Politoscope a aussi passé au crible l'activité sur X avant et pendant les élections législatives de 2024 (les 29 et 30 juin, pour le premier tour, avec un second tour les 6 et 7 juillet). « Notre étude [1] a permis de caractériser un processus d'affaiblissement puis d'inversion du front républicain à l'approche des législatives de 2024 et d'identifier les stratégies de subversion qui l'ont favorisé », explique le chercheur.

Comme David Chavalarias a pu le constater : « La diffusion d'information par le design algorithmique de la plateforme est biaisée sur des dimensions qui sont nuisibles au débat public, notamment la dimension de la toxicité ». C'est-à-dire que les messages qui sont toxiques, qui comportent des insultes, des invectives… sont sur-recommandés aux utilisateurs. Dans les campagnes électorales, ce biais amplifie les contenus à scandale, la polémique, la dénonciation, par rapport à des propos plus neutres ou constructifs. Autre problème, des contenus provenant de faux sites, de répliques de sites gouvernementaux et d'actualités, avec des messages visant par exemple à retirer totalement le soutien de la France à l'Ukraine, à dénigrer le gouvernement en place, et notamment en France, le président Macron, ont aussi été repris et re-tweetés.

INGÉRENCE ÉTRANGÈRE

Profitant de cette dynamique et des réglages de la plateforme, « des communautés numériques minoritaires en termes de nombres de comptes observés — 60 000 comptes environ, soit 0,4% des millions de comptes étudiés — parviennent à être majoritaires sur les grands thèmes clivants qui fracturent la société française, et à en amplifier plusieurs à la fois. Il y a une organisation sur les réseaux sociaux d'un leadership communautaire pour entraîner un certain nombre de personnes dans ce genre de thématiques ». Parmi ces thématiques très polarisantes pour la société française et faisant l'objet de fortes polémiques — sur lesquelles ces communautés se sont mobilisées systématiquement — figurent le soutien militaire de la France à l'Ukraine, le conflit Israël–Gaza, l'islamo-gauchisme, le wokisme, le dénialisme climatique, les anti-vaccins, etc.

L'effet « boule de neige » est aussi recherché, par un cumul des thèmes de divisions. « Les contenus et retweets montrent aussi des stratégies d'amalgames, d'affaiblissement et de discrédit des partis et leaders politiques gouvernementaux et de l'extrême-gauche, favorisant des discours de "décohésion" de la société française », note le spécialiste. L'étude du Politoscope montre que de tels contenus ont en effet été repris en priorité et re-tweetés. Pour David Chavalarias, ces opérations servent bien des stratégies d'influence, via de faux comptes, des fermes de trolls, de l'*astroturfing*, et des soutiens politiques en France. L'analyse des activités en ligne et des cartographies de contenus et retweets retrace une convergence d'intérêts entre le régime de Poutine, l'extrême-droite française et des communautés de comptes, visant à affaiblir les efforts de coalition des partis traditionnels au second tour des législatives et à compromettre tout barrage contre l'extrême-droite, présentée comme banalisée.

À la lumière de ces résultats, le chercheur du CNRS émet plusieurs recommandations. Elles s'adressent autant aux autorités et institutions, qu'au public, notamment en matière de régulation européenne des espaces numériques, de développement de plateformes alternatives plus éthiques, et d'éducation aux médias (voir chapitre 6).

LE SAVIEZ-VOUS ?

CARTOGRAPHIES ET OUTILS DE DÉTECTION

Dans le cadre du projet Politoscope, les chercheurs ont développé des méthodes mathématiques et informatiques pour permettre une représentation de l'évolution des tendances des partis politiques, de leurs alliances et de leurs discours à partir des informations disponibles en ligne et ce, quasiment en temps réel. De plus, des outils de détection de comportements collectifs anormaux au niveau des groupes ont été créés. Ces derniers étant souvent la signature de manipulations massives : faux comptes, fermes de trolls, robots, astroturfing. Sur cette base, le Politoscope recueille et analyse un ensemble de données considérable : entre 500 000 et deux millions de tweets par jour, comprenant des messages politiques. Une cartographie de la structure globale, calculée numériquement, permet de refléter les proximités idéologiques des comptes analysés, les flux d'échanges et des retweets, par principaux thèmes, et de repérer des phénomènes d'amplification. Les études du Politoscope se donnent pour mission de « mieux comprendre comment se forment les groupes sociaux dans les environnements numériques, étudier les évolutions du militantisme politique et caractériser les stratégies des partis politiques et les manipulations d'opinion », témoigne David Chavalarias. Le but étant d'arriver ensuite à anticiper, alerter et sensibiliser les citoyens à ces thématiques.

CHAPITRE 4

Inde
Microciblage, chatbots, « fantômes » et avatars

L'Inde représente un très gros marché pour des entreprises proposant du matériel de campagne politique agressif basé sur l'IA. Ce secteur en plein essor profite d'un vide juridique et s'engouffre dans des brèches parfois inattendues. Analyse avec Anushka Jain, du Digital Futures Lab.

Basé à Goa, le Digital Futures Lab [1] est un centre de recherche multidisciplinaire qui s'intéresse aux interactions complexes entre technologies et société dans les pays du sud. Juriste et spécialiste des dimensions éthiques concernant les applications de l'IA, Anushka Jain y travaille comme chercheuse associée : « Je m'intéresse à la technologie et à ses réglementations, avec une spécialisation concernant la gestion des données personnelles en Inde, notamment par des outils de surveillance recourant à l'IA. J'ai beaucoup étudié les technologies de reconnaissance faciale, les biais discriminants et les impacts sur les droits humains. En particulier la dimension du genre qui fait l'objet de fortes divisions en Inde ».

LE RECOURS AUX NOUVELLES APPLICATIONS D'IA GÉNÉRATIVE EST DEVENU MASSIF

Le Digital Futures Lab suit aussi de près les problèmes actuels de désinformation : le recours aux nouvelles applications de l'IA générative est en effet devenu massif dans ce pays dirigé depuis dix ans par un politicien ultra-conservateur nationaliste et religieux, le Premier ministre Narendra Modi — à la tête du Bharatiya Janata Party, parti d'extrême droite nationaliste hindou. Les législatives indiennes ont eu lieu du 19 avril au 1er juin 2024, mobilisant près de 970 millions d'électeurs — soit un des plus grands scrutins de l'Histoire. Au-delà des exemples concrets de manipulations (voir encadré), quelles leçons tirer de cette période mouvementée d'élections ?

ZONES GRISES

Beaucoup d'attention a d'abord été portée sur l'usage de *deepfakes*, à des fins de perturbation et de désinformation. « Or, après les élections, plusieurs études ont conclu que certes, tout cela avait eu un impact et que l'IA avait été abondamment utilisée en ce sens, mais que finalement, le processus électoral n'en avait pas autant été affecté », témoigne l'experte. D'autres utilisations plus inattendues de l'IA se sont popularisées. Par exemple, la capacité des outils d'IA à « ressusciter » un politicien décédé, avant les élections. Une « version synthétique » de l'ancienne vedette de cinéma et leader politique M. Karunanidhi, mort en 2018, s'est ainsi adressée aux électeurs dans un discours de huit minutes. Le même subterfuge a été mis en scène en Indonésie (voir encadré).

Là aussi, le vide juridique et l'absence de réglementations profitent aux divers acteurs. « Ces pratiques s'avèrent très problématiques, comme le fait de recourir à l'IA pour attribuer des déclarations à des personnes qui n'en sont pas les auteurs », relève Anushka Jain. Ces brèches dans lesquelles s'engouffrent des producteurs et diffuseurs de contenus synthétiques visant à influencer l'opinion politique lui semblent urgentes à étudier, avec leurs enjeux éthiques, au-delà « des *fake news* ou des *deepfakes*, où le caractère mensonger et l'intention de manipulation sont bien plus clairs ».

Elle souligne un autre aspect qui constitue aussi « une sorte de zone grise » en termes de réglementation : la « guerre des mèmes » à laquelle se sont livrés les différents candidats et leurs équipes de campagne. Ce type d'intervention se fonde sur une « culture de l'Internet » et un statut très ambigu des images fabriquées qui doivent interpeller, signale la chercheuse : « Il n'y a rien d'illégal à créer des mèmes et des parodies avec de l'IA générative en période d'élection, et ils ont été très nombreux. Mais comme nous l'avons remarqué, beaucoup comportaient une part de dénigrement, étaient utilisés pour cibler spécifiquement certains politiciens… Que ce soit en Asie ou aux États-Unis, pourquoi telle ou telle image détournée devient-elle populaire, comment se retrouve-t-elle diffusée et rediffusée, devient-elle importante ou pas ? »

Dans le cas des législatives indiennes de 2024, il est estimé que ce marché spécifique axé sur la création de contenus génératifs pour nourrir et amplifier des campagnes déstabilisantes a atteint 15 à 16 millions de dollars, et seulement en prenant en compte les demandes des partis politiques. Un responsable a avoué que sa société comptait sur un revenu de 241 000 dollars en six semaines, pour ces prestations. Anushka Jain mentionne une société en particulier qui se targue d'être un leader dans le domaine : The Indian Deepfaker, spécialisée dans la production de médias et contenus synthétiques, entre autres pour des chaînes de télévision (comme Shine TV, au Royaume-Uni), l'industrie du cinéma et Netflix. Leur promesse marketing : « repousser les limites du *storytelling* avec la puissance des technologies de *deepfake* basés sur l'IA ».

La juriste rappelle aussi le contexte économique singulier de l'Inde, où le développement et l'innovation digitale en général sont fortement encouragés, y compris par le gouvernement et les institutions. « Ce secteur est perçu comme un des piliers de la société indienne. Aux yeux de tous, il est très important pour l'avenir du pays. Je ne vois pas les politiciens renoncer à utiliser l'IA ».

LE SAVIEZ-VOUS ?

EXEMPLES CONCRETS

Avertissements aux partis politiques. Avant juin 2024, une vidéo *deepfake* montrant le célèbre acteur de Bollywood, Aamir Khan, faisant la promotion d'un parti politique, est devenue virale, avant d'être démasquée et dénoncée à la police de Mumbai. Plusieurs autres vidéos *deepfake* ont finalement conduit la Commission électorale de l'Inde à envoyer aux partis politiques un courrier d'avertissement leur demandant de respecter un code de conduite, de ne pas recourir à de tels subterfuges, et leur donnant un délai de trois heures pour retirer les contenus visés.

Chatbots et avatars. Le Ministère des technologies de l'électronique et de l'information serait une autre entité concernée par la régulation, ajoute la juriste, tandis que les usages de vidéos synthétisées, d'avatars et de *chatbots* servant au micro-ciblage des électeurs se sont largement répandus. Un exemple : le démarchage téléphonique via des robots. Quinze millions d'appels générés par l'IA ont touché les électeurs en l'espace de deux mois, venant « brouiller les lignes ».

Multilinguisme. Un autre exemple de performance réalisée grâce à l'IA a attiré l'attention : une vidéo de Narendra Modi s'adressant aux électeurs en de multiples langues, alors qu'il ne les parle pas. Les progrès de la traduction automatisée en temps réel ont fait une forte impression en Inde, pays qui comprend plus de 80 langues officielles. La mode des « résurrections » synthétisées de personnes charismatiques défuntes s'est aussi très vite développée en Asie comme un moyen d'influence très efficace, et très prisé.

EN INDONÉSIE

En Indonésie, un faux général Suharto (mort en 2008) généré par IA est soudain apparu dans les médias et sur les réseaux sociaux, s'adressant aux électeurs dans un discours de trois minutes, peu avant les élections présidentielles et parlementaires du 14 février 2024, pour les encourager à aller voter dans le camp des traditionalistes. Cette vidéo d'abord publiée sur X par un député du parti Golkar — fondé en 1964 par l'armée, et instrument du régime autoritaire — a rapidement atteint 4,7 millions de vues. Elle a été re-postée sur TikTok, Facebook, YouTube. L'équipe du président Prabowo Subianto, ancien militaire accusé de nombreux méfaits, a aussi recouru à de multiples usages de l'IA pour améliorer son image, paraître rajeuni, plus sympathique ; et même, créer un avatar sous forme de personnage de dessin animé, jovial et attendrissant, ciblant les jeunes électeurs ne connaissant pas son lourd passé.

CHAPITRE 5

Fact checking et journalisme collaboratif

Depuis une dizaine d'années, plusieurs modèles de journalisme participatif et citoyen ont vu le jour pour réagir à la montée en puissance des fausses nouvelles et opérations de désinformation. La mutualisation des ressources, outils et compétences reste une voie à explorer, aussi en appui aux équipes de fact checking des rédactions.

En France, les cellules de *fact checking* de nombreuses rédactions sont mises à rude épreuve. Comment faire face à la profusion de fausses nouvelles et d'intox amplifiées sur les réseaux sociaux ? Aux périodes de crise propices à la forte diffusion de propos complotistes — comme durant la pandémie de Covid-19 ? Ou à la quantité et qualité des *deepfakes* en période d'élections ? La mise en réseau de rédactions et d'associations de journalistes professionnels dédiés à l'investigation et à la vérification d'informations peut contribuer à pallier ce problème (voir l'encadré). D'autres pistes sont aussi explorées.

Citizen Facts a ainsi été imaginé par deux professionnels des médias, basés à Paris. Journaliste et formatrice, Aude Favre a travaillé pour des émissions de télévision axées sur l'investigation. Un premier projet émerge en 2015 après les attentats contre Charlie Hebdo. Avec son collègue Sylvain Louvet, elle lance une association de journalistes engagés contre la désinformation de masse [1]. Elle compte près de cinquante journalistes aujourd'hui, répartis dans toute la France. Aude Favre poursuit ce type d'expériences en créant ensuite sa chaîne Youtube WTFake, et propose lors du premier confinement en 2020 d'enquêter avec elle sur les *fake news* et les théories du complot. D'autres enquêtes du même type sont mises en place. Ex-directeur éditorial chez Brut et réalisateur de films d'investigation, Sylvain Louvet a entre autres signé avec elle un documentaire sur le business de la désinformation et le marché publicitaire sur Internet [2]. Comme il le rappelle : « Nous avons testé le modèle de la rédaction collaborative et posté un appel à témoins sur notre plateforme. Une cinquantaine de personnes ont participé à l'expérience. Pendant deux mois, des citoyens ont parcouru des sites complotistes pour identifier quelles publicités s'affichaient, et retrouver les sources de financements. Ils nous ont aidés à mener une enquête minutieuse sur l'affichage de la publicité programmatique, les Google ads et les encarts publicitaires ». En 2023, les deux journalistes ont concrétisé avec Arte une série documentaire basée sur ce même principe, en formant une rédaction citoyenne constituée de plus d'un millier de personnes à travers l'Europe [3].

PENDANT DEUX MOIS, DES CITOYENS ONT PARCOURU DES SITES COMPLOTISTES POUR IDENTIFIER LES PUBLICITÉS ET SOURCES DE FINANCEMENTS

DÉFIS DU COLLABORATIF

La perspective du journalisme collaboratif pose de nombreux défis. Il faut entre autres trouver des moyens sécurisés de communication, de travail et d'analyse des données adéquats. Mais aussi, réunir une diversité de compétences parmi des bénévoles, diriger et coordonner une équipe participative. Enfin, se concerter sur la mise en forme, en valeur et la diffusion des investigations réalisées. « Au-delà du bénévolat, si nous voulions conserver ces enquêteurs-citoyens, il nous fallait aussi trouver un nouveau modèle. Avec la *blockchain*, nous avons eu cette idée de traçabilité des contenus. La personne qui déposera des informations et documents sur la *blockchain* sera pour nous identifiable, ce qui favorisera une manière plus vertueuse de collaborer », explique Sylvain Louvet. Le développement de ce projet par la création d'une nouvelle plateforme, l'étude d'un nouveau mode de gouvernance, d'un système de fidélisation et rémunération des contributeurs, est soutenu depuis octobre 2024 par un programme de l'IMI, avec la RTS et différentes universités [4]. À court terme, le projet Citizen Facts devrait proposer une offre de services indépendante de *fact checking* et d'enquêtes, en partenariat avec des médias européens.

UN MODÈLE INSPIRANT

Fort de cinq ans d'activités et d'expériences, The Markup (voir chapitre 2) représente un modèle inspirant pour le développement d'un journalisme participatif et citoyen, autant comme moyen de contrer la désinformation que pour contribuer à l'éducation aux médias. Rédaction collaborative à but non lucratif, The Markup se définit comme « un nouveau type d'organisation médiatique », réunissant des profils et métiers complémentaires. Sa co-fondatrice, Sisi Wei, rappelle son mode de fonctionnement particulier, sa spécialisation et les atouts qui lui ont permis de mieux se profiler dans le paysage médiatique :

➜ **Une complémentarité de l'équipe**, composée de journalistes, analystes, programmeurs, ingénieurs. L'idée étant de mobiliser toutes leurs compétences pour mener des investigations, en groupe ou par tandems.

➜ **Une démarche participative**. The Markup lance des appels à collaborer et reçoit des candidats de tous profils, intéressés à apprendre à faire du journalisme mais aussi à utiliser de nouveaux outils, ou ayant une expertise préalable d'applications d'IA. « Ce modèle collaboratif et citoyen fonctionne bien », constate Sisi Wei, « beaucoup de collaborateurs bénévoles, d'informations et de données sont proposées et postées ».

➜ **Un choix éditorial de sujets partant de préoccupations concrètes exprimées par des communautés**. Il s'agit de se focaliser sur des investigations utiles à la société, sur des sujets d'intérêt public, notamment le rapport à la technologie (protection des données, accès ou non aux données et informations gouvernementales, des institutions, etc.).

➜ **Un retour aux acteurs**. The Markup offre aux communautés d'utiliser ces enquêtes pour mieux défendre leur cause et faire évoluer les réglementations. « Avec ces enquêtes, les participants ont ainsi le sentiment de pouvoir contribuer à changer quelque chose, dans la société, qu'ils ont la capacité à faire la différence, en tant que citoyens », résume Sisi Wei. Il y a un effet d'impact, de proximité avec le sujet d'investigation proposé.

➜ **Un modèle de datajournalisme**. Axé sur le recueil, le traitement et la visualisation de données.

➜ **Une approche transparente et didactique**, autant sur la manière d'enquêter, de documenter, ou d'utiliser des outils de collecte et d'analyse de données.

LE SAVIEZ-VOUS ?

Depuis 2015, plusieurs médias francophones ont porté le projet de la plateforme « Source sûre » [5], avec un contact sécurisé entre journalistes et lanceurs d'alerte. D'autres outils sont recommandés par le Global Investigative Journalism Network, qui a mis en ligne en 2024 un « Guide à l'usage des journalistes d'investigation en période électorale » [6].

Ces précieuses initiatives de mise en réseau sont-elles trop « occidentalo-centrées » ? Elles gagneraient à mieux tenir compte de la très grande disparité des contextes. C'est ce que souligne à Lausanne Caroline Vuillemin, directrice de la Fondation Hirondelle [7]. Cette organisation suisse à but non lucratif a depuis la fin des années 1990 créé et développé 23 médias d'information — notamment des radios FM — dans 22 pays confrontés à des crises majeures. Les régions d'Afrique francophone dans lesquelles la Fondation Hirondelle intervient, comme au Mali, Niger, Burkina Faso, en République démocratique du Congo et en République centrafricaine, sont exposés de manière très différente aux risques de désinformation en période électorale. Une bonne partie de la population n'a pas accès à Internet, ni à une couverture du réseau. « L'essentiel des manipulations continue à se faire par le bouche-à-oreille et par les médias traditionnels, des radios à la solde de divers intérêts, ou de radios communautaires qui n'ont pas les outils de vérification nécessaires, ni les bonnes pratiques journalistiques ». La messagerie WhatsApp est aussi un canal de diffusion privilégié. Enfin, le rapport aux *fake news* et aux possibilités de *debunking* se pose très différemment « dans certains pays, où le premier acteur de la désinformation… c'est l'État », conclut Caroline Vuillemin.

CHAPITRE 6

Pistes et recommandations

Chercheurs, experts et professionnels des médias ont suggéré en 2024 plusieurs mesures sur le court, moyen et long terme. Ces dernières visent à mieux définir les parts de responsabilité des acteurs en présence, réguler les usages de l'IA et limiter les impacts de stratégies de désinformation sur les processus électoraux. Tour d'horizon.

CENTRES DE RECHERCHE

→ Dresser un état des lieux de types de risques et catégories d'impacts réels de la diffusion de désinformation sur les plateformes en période électorale, en évitant de les surévaluer (voir les études [1, 2]).

→ Analyser et mieux identifier les risques systémiques des plateformes et de leur *black box* (par ex., nature et effets des filtres algorithmiques, et leur possible exploitation pour des opérations de désinformation).

→ Développer des outils qui permettent aux utilisateurs de plateformes de mieux identifier leurs interlocuteurs ; une fonction qui permette de ne faire apparaître que les messages de vraies personnes et non ceux produits par des IA ; favoriser une certification / labellisation qui permette d'identifier la nationalité et le lieu de diffusion des interlocuteurs.

→ Encourager avec les développeurs des opportunités d'applications positives dans le domaine de la politique, par exemple la création d'outils basés sur l'IA permettant aux électeurs : d'être mieux informés sur les candidats, de pouvoir les évaluer, ainsi que leurs engagements ; de générer des synthèses sur les enjeux et programmes politiques, accessibles au public ; de faciliter une communication impartiale et non biaisée avec les milieux politiques ; et pour les petits partis et les représentants des minorités, d'égaliser les règles du jeu lors du processus électoral.

→ Favoriser les initiatives portant sur le *debunking*, pour que les électeurs mettent mieux en doute les informations politiques sensationnelles ou scandaleuses, surtout lorsqu'elles sont publiées à l'approche du jour du scrutin (voir le manuel [3])

→ Étudier comment modifier le type de scrutin, adapter les systèmes et processus électoraux (en s'inspirant par exemple du vote à second tour instantané, aussi appelé vote alternatif), pour limiter le ciblage, la propagation de scandales et *deepfakes* sur des candidats particuliers à un moment politique décisif.

INSTITUTIONS

➔ Améliorer la régulation des plateformes, en appliquant le Digital Service Act et le EU AI Act, et en les intégrant et adaptant dans de nouvelles réglementations au niveau national.

➔ Promouvoir des lois et initiatives visant à limiter la concentration et la centralisation de l'information par les plateformes et des entreprises technologiques développant les outils d'IA générative.

➔ Encourager la veille et le monitoring d'opérations de désinformation lors de processus électoraux (exemples : l'Observatoire européen des médias numériques [4] ; le Centre de suivi de la mésinformation sur les élections européennes, par NewsGuard [5]).

➔ Définir quels types de plateformes soutenir ou créer pour un espace numérique européen sécurisé, entre autres face aux stratégies de déstabilisation opérées depuis l'étranger.

➔ Développer des plateformes alternatives qui soient propices au débat public et notamment, des réseaux sociaux européens ne présentant pas des biais algorithmiques posant problème (comme l'amplification des *fake news* et de *hate speech*), respectant une certaine charte éthique et les données personnelles des utilisateurs (par ex., le projet Mastodon [6]).

➔ Favoriser, sur les plateformes, la certification et la réputation de l'identité digitale des utilisateurs.

ENTREPRISES TECH

➔ Tracer et labelliser les contenus générés par l'IA, notamment les images (*watermarking, labeling, signing*).

➔ Renforcer les démarches d'*accountability* et de traçabilité, pour les diffuseurs de contenus (exemple au niveau gouvernemental : à Taïwan, système de licences accordées aux médias, avec marques de certification).

➔ Surveiller l'utilisation des *chatbots* pour le micro-ciblage des votants, les contenus trompeurs destinés à dissuader la participation, les *deepfakes* et l'écosystème de l'information, en particulier la diffusion non sourcée d'informations sur les processus électoraux.

➔ S'assurer que les *chatbots* dirigent les utilisateurs vers des informations provenant de sources officielles, notamment sur les questions de savoir comment et où voter.

➔ Encourager, sur tous ces points, un consensus sur les approches et les technologies à développer et sur les standards globaux à adopter.

➔ Soutenir des projets collaboratifs portant sur le *debunking* (par ex. PolitiFact, lancé en 2007 déjà par Poynter Institute for Media Studies [7]).

➔ Promouvoir l'éducation aux médias et aux usages de l'IA, notamment face aux *fake news* et aux opérations de désinformation en période électorale.

MILIEUX POLITIQUES

➔ S'engager publiquement à ne pas utiliser de contenu trompeur, déstabilisant et agressif généré par l'IA.

➔ Respecter les réglementations en vigueur en la matière (aux États-Unis, le recours aux *deepfakes* en période électorale est déjà interdit ou limité dans 19 États ; en France, une loi de mai 2024 vise à sécuriser et à réguler l'espace numérique et sanctionne l'utilisation détournée des images de politiciens dans des *deepfakes* [8].

➔ Renoncer au microciblage et aux *chatbots* en période électorale, et au démarchage téléphonique automatisé (aux États-Unis, ce type d'appel recourant à des voix synthétiques est interdit et sanctionné depuis février 2024 par la Federal Communications Commission, voir article [9]).

➔ Favoriser les débats dans les milieux politiques sur ces questions et trouver un consensus inter-partis sur les règles globales à respecter, dans une charte commune ; notamment à la lumière d'exemples de recours à l'IA en Suisse dans le cadre de campagnes politiques, sur des cas déjà dénoncés ou sanctionnés [10, 11, 12, 13].

MÉDIAS

➔ Décourager les opérations de désinformation et de manipulation en évitant de couvrir des sujets dont le seul argument en faveur de l'intérêt médiatique est l'utilisation de contenu généré par l'IA.

➔ Inciter à l'éducation aux médias et à l'IA grâce à une couverture non sensationnaliste du rôle de l'IA dans les élections.

➔ Réfléchir en commun, avec les différentes instances concernées et la société civile, à la manière de couvrir et évaluer rapidement les contenus potentiellement trompeurs, et à détecter les *deepfakes*.

➔ Mutualister les projets et ressources pouvant servir à l'éducation aux médias et à l'IA (par ex., The Wired AI Elections Project [14] ; voir chapitre 5).

À propos
L'Initiative for Media Innovation (IMI) est un consortium dédié à la création de synergies entre le monde académique et les acteurs des médias et de l'information. Cette alliance unique rassemble l'excellence scientifique et journalistique pour encourager la recherche interdisciplinaire et proposer des solutions transformatrices. Ensemble, les membres de l'IMI œuvrent à garantir une information de qualité, fiable et accessible à tous.

Membres fondateurs

EPFL · SRG SSR · Triangle Azur | Université de Genève, Université de Lausanne, Université de Neuchâtel

Membre associé

Hes·so VALAIS WALLIS

Références

Novembre 2024

Recherches, entretiens et rédaction : Gilles Labarthe (PhD), en collaboration avec Mounir Krichane, Christophe Schenk et Julie Schübpach, IMI

Graphisme : Anne-Sylvie Borter, Centre d'impression EPFL

Photos : Jr Korpa, Ahmad Odeh, Simon King, Nonsap Visual, Nikunj Gupta, Sebastian Hages et Nelson Wong sur Unsplash

Impression au Centre d'impression EPFL, certifié myClimate, sur papier 100% recyclé