LINC
CNIL.

DOSSIER

Elections : quelles influence de l'IA sur notre vote ?

Juillet 2024

linc.cnil.fr

Alexis Leautier, Ingénieur IA

Alors que la qualité de l'information en ligne est préoccupante depuis plusieurs années, en particulier dans un contexte électoral, la démocratisation des outils d'IA pourrait être la cause d'évolutions inédites. De nombreuses initiatives législatives visent à limiter ce risque qui pèse sur le fonctionnement démocratique de notre société, mais leur analyse laisse entrevoir certaines failles.

Six ans après les révélations du scandale Cambridge Analytica dénonçant entre autres le rôle de Facebook dans le référendum ayant conduit au Brexit, l'influence des réseaux sociaux et des moteurs de recherche sur les processus électoraux suscite de nouvelles interrogations. La démocratisation des outils d'IA générative et leur intégration grandissante dans les moteurs de recherche et les réseaux sociaux pourrait ouvrir la porte à une phase de désinformation d'une ampleur inédite tant les productions de ces outils paraissent aujourd'hui réalistes. Les systèmes de recommandation et de ciblage, dont les risques sont bien identifiés aujourd'hui mais encore mal maîtrisés, pourraient servir de tremplin à ces productions et ainsi propager hypertrucages et hallucinations grâce au profilage des utilisateurs et à l'exploitation de leurs données personnelles.

Alors qu'un quart de la population mondiale devrait se rendre aux urnes en 2024, les législateurs de certains Etats semble se saisir du sujet à bras-le-corps. En Europe, le règlement relatif à un marché unique des services numériques (*Digital Services Act*), surnommé RSN, déjà en application, contient des règles relatives aux systèmes de recommandation et au ciblage publicitaire, ainsi que des règles relatives à l'évaluation et l'atténuation des risques systémiques pour les processus électoraux pouvant être causés par les services des très grandes plateformes et très grands moteurs de recherche. En vue des élections européennes de juin 2024, la Commission européenne a ainsi publié ses lignes directrices pour les très grandes plateformes et très grands moteurs de recherche sur l'atténuation des risques systémiques pour les processus électoraux en application du RSN. Ces lignes directrices tiennent compte par anticipation de certaines des règles prévues par le règlement relatif à la transparence et au ciblage de la publicité à caractère politique, qui entrera pleinement en application à l'automne 2025, ainsi que par le règlement établissant des règles harmonisées pour l'intelligence artificielle, ou « règlement IA » (*AI Act*), qui vise particulièrement l'IA générative et les systèmes visant à influencer les résultats d'une élection. Au niveau national cette fois-ci, la loi du 21 mai 2024 visant à sécuriser et à réguler l'espace numérique (dite loi SREN) prévoit elle aussi des dispositions relatives à la publication de contenus génératifs. Enfin, une loi sur les ingérences étrangères en France, pourrait prochainement venir compléter ce tableau en prévoyant notamment l'utilisation de l'IA pour la détection de tentatives d'ingérences.

Bien que cet arsenal puisse sembler suffisant pour anticiper et contrôler les risques liés à l'utilisation de l'IA dans un contexte électoral, son application en pratique questionne, comme en témoignent les enquêtes initiées par la Commission Européenne à l'encontre de Meta et Microsoft. Si le respect des textes par ces éditeurs de services est une chose que les enquêtes de la Commission permettra de vérifier, l'ambition et la faisabilité pratique des mesures prévues en est une autre. La cartographie des risques, obligations légales et

solutions techniques présentée dans ces articles a pour ambition d'apporter des réponses à cette seconde interrogation.

En explorant trois utilisations spécifiques de l'IA, le Linc vous propose d'évaluer les potentiels impacts de l'intelligence artificielle sur les processus électoraux, à commencer par l'utilisation de l'IA générative.

- Le discours bien réel des contenus artificiels,
- Des annonceurs aux utilisateurs : un cheminement algorithmique,
- L'autocontrôle, un nouvel apprentissage pour l'IA.

Si ces utilisations de l'IA sont principalement explorées ici sous l'angle des risques qu'ils comportent, les utilisations pouvant au contraire bénéficier aux processus démocratiques ont été explorées dans le cahier IP n°7 « *Civic tech*, données et *Demos* ». Si le constat d'un désenchantement depuis un idéal d'internet comme espace public revitalisé y est partagé, des remèdes liés au déploiement de civic tech, à internet comme lieu de contre-pouvoir et de mobilisation y sont également mis en avant.

SOMMAIRE DU DOSSIER

Le discours bien réel des contenus artificiels [1/3]

C'est tout d'abord par la création de contenus que l'IA – ici générative – pourrait contribuer à influencer le vote. Qu'il s'agisse d'un chatbot utilisé par les électeurs pour s'informer sur une élection, de matériel de campagne généré par un parti pour communiquer sur son programme, ou encore d'hypertrucages visant à manipuler l'opinion publique, les outils d'IA générative pourraient rendre plus facile la communication tant d'information que de désinformation.

Source : [Pexels](#)

La génération de contenus officiels pour informer les électeurs

Deux situations doivent être distinguées en ce qui concerne l'utilisation de l'IA générative pour informer les électeurs.

1. L'utilisation des chatbots généralistes

Tout d'abord, les chatbots grand public tel que ChatGPT, Le Chat, Gemini, ou encore Bing Chat, peuvent être interrogés par leurs utilisateurs sur la tenue des élections. Dans ce premier cas, l'outil n'a pas été conçu ou adapté pour répondre à ces demandes spécifiques. En raison des limitations portant sur le corpus d'entraînement des grands modèles de langage sous-jacents, et du comportement naturellement probabiliste de ceux-ci, les

réponses fournies par ces outils risquent fortement de ne pas être correctes ou pertinentes. C'est ce que soulève Democracy Reporting International dans un rapport publié en avril 2024. Après avoir interrogé plusieurs outils en mars 2024, l'organisation a conclu que les informations fournies étaient fréquemment inventées, au point de parfois fournir des dates erronées pour la tenue d'un scrutin. En pratique, ces erreurs peuvent se manifester de plusieurs manières :

- **les hallucinations**, c'est-à-dire la production d'informations ne provenant d'aucune source en raison du fonctionnement probabiliste de l'IA générative ;
- **l'incomplétude des réponses**, comme par exemple un parti qui ne serait pas listé dans la réponse à la question « pour qui puis-je voter ? »,
- **l'inexactitude des réponses**, qui diffère des hallucinations par exemple lorsque l'information fournie est bien sourcée, mais qu'elle est obsolète ou qu'elle a été, volontairement (par un empoisonnement des données comme décrit dans un article du Linc sur les attaques en IA) ou non, corrompue. Ce risque a notamment été pointé par Newsguard qui a relevé que certains chatbots reproduisaient des informations provenant de sites de désinformation russe,
- **les biais dans les réponses, ou dans la qualité des réponses**, ce qui peut entraîner une moins bonne information de certains groupes de personnes (en raison de leur langue, moins maîtrisée par le système par exemple), ou encore désavantager certains partis politiques.

Les conséquences de ces erreurs pourraient être de plusieurs ordres :

- **Une diminution de la participation**, par :
  - une désinformation telle que certains électeurs ne seraient pas en mesure de voter (en se trompant sur la date butoir d'inscription sur les listes électorales par exemple),
  - une confusion dans les réponses ou une friction dans l'utilisation des outils entraînant une démobilisation des électeurs (renvoi vers des sources d'informations contradictoires ou dépassées, information incomplète, etc.),
- **Une influence sur le résultat des élections**, par :
  - La démobilisation de certains groupes parmi la population, en raison par exemple du plus grand taux d'hallucinations des outils dans certaines langues.

Ce type d'utilisation des outils par les électeurs étant inévitable, d'autant plus qu'elle témoigne d'un intérêt pour les élections, des mesures devraient être prises afin d'éviter la désinformation. De telles mesures existent et incluent :

- **L'alignement des modèles**, qui inclut des techniques telles que le *reinforcement learning from human feedback* (RLHF), la *direct preference optimization* (DPO), ou encore l'IA « constitutionnelle ». Ces techniques ne sont toutefois pas infaillibles et peuvent être contournées (par des techniques d'injection de prompt comme le « *jailbreak* » de ChatGPT surnommé DAN par exemple).
- **L'utilisation de filtres sur les prompts fournis en entrée**. Cette solution, qui repose généralement sur la classification du prompt avant de générer une réponse, permet d'identifier les demandes auxquelles il est préférable de ne pas répondre.
- **L'utilisation de filtres sur les réponses**, qui a l'avantage de porter directement sur la réponse fournie à l'utilisateur.

Lorsque les prompts fournis en entrée se rapportent à un sujet sur lequel les hallucinations peuvent entraîner un risque trop important, deux stratégies peuvent être prises :

- fournir une réponse type indiquant que l'outil n'est pas habilité à répondre ou renvoyant vers un site officiel (comme le propose l'entreprise Anthropic sur Claude par exemple),
- répondre à la question utilisant une technique plus robuste comme le *retrieval augmented generation* (RAG).

Cette deuxième solution permet de ne pas renoncer à l'information des électeurs via l'outil, mais elle reste sujette à la génération d'hallucination. Dans ce cas, davantage de précautions pourraient être prises pour informer l'utilisateur sur le fonctionnement, pour limiter le texte généré et plutôt favoriser les copier-coller dans les réponses, ou encore pour indiquer les sources d'où proviennent les informations.

2. Vers la création de chatbots dédiés

Dans un second scenario, des outils pourraient être conçus dans l'objectif spécifique de fournir des informations aux électeurs sur la tenue des élections. Bien qu'aucun de ces outils n'ait été observé à ce jour, ce scénario semble néanmoins probable étant donné le rôle croissant que les chatbots seront amené à tenir pour informer et éduquer. Ces outils pourraient avoir un rôle bénéfique en apportant des informations adaptées aux connaissances de l'utilisateur, et en les reformulant lorsqu'elles manquent de clarté. Toutefois, plusieurs risques pourraient persister à un niveau important malgré les mesures prises, questionnant l'intérêt final d'un tel outil. Comme dans le cas des IA généralistes, ces risques comprennent notamment la génération d'hallucinations, de réponses incomplètes, inexactes, ou biaisées.

La complexité du fonctionnement des élections de grande ampleur, comme les élections européennes en particulier, rend la réponse à des questions d'apparence simples bien plus difficile qu'anticipé. Une question telle que « où puis-je voter ? » peut en effet trouver des réponses différentes selon la nationalité d'après le site elections.europa.eu : le pays de résidence, selon si ce dernier appartient à l'UE ou non, puis selon la commune d'inscription aux listes électorales, qui peut être « la commune de votre résidence principale ou du lieu où vous habitez depuis au moins 6 mois ; la commune où vous êtes assujetti.e aux impôts locaux depuis au moins 2 ans ; la commune où la société (dont vous êtes le.la gérant.e ou l'associé.e majoritaire ou unique) est inscrite depuis au moins 2 ans ; pour les fonctionnaires, la commune de votre résidence d'affectation obligatoire ». Après ce cheminement, d'après le site du ministère de l'intérieur, la réponse finalement utile à l'utilisateur (c'est-à-dire le lieu physique où il peut se rendre à l'isoloir) ne pourra être trouvée que via … un téléservice dédié, ou sur sa carte électorale.

D'une manière générale, les informations relatives à un vote pouvant changer (en raison d'un événement, de l'indisponibilité du bâtiment prévu pour le vote, d'un changement dans les listes, etc.), la conception d'un outil dédié peut sembler être un pari risqué si celui-ci ne se réfère pas de manière dynamique à une source d'informations tierce fiable. C'est justement le sens de la recommandation de l'organisation américaine Democracy Works qui

déconseille d'entraîner une IA générative sur son corpus d'informations sur les élections américaines. L'approche de Democracy Works consiste plutôt à utiliser l'IA générative pour comprendre la question de l'utilisateur, requêter l'information correspondante dans une base de données tenue et mise à jour par l'organisation, et de fournir cette information telle quelle pour ne pas en dénaturer le fond. Ainsi, l'IA peut fournir des informations actualisées et en citer la source. Democracy Works recommande également que cette mise en œuvre s'accompagne de nombreuses mesures de sécurité préalables, s'effectue après un certain nombre de tests, et sous réserve d'audits réguliers a posteriori.

La génération de contenus pour la production de matériel de campagne et la prospection

Si l'utilisation de l'IA générative par les partis pour faciliter leur campagne peut sembler anecdotique pour certains usages – comme pour retoucher les affiches de la candidate Juliette de Causans – à une plus grande échelle, cette utilisation peut avoir des effets (désirés ou non) notables. En effet, l'IA générative peut être utilisée dans la génération de matériel de campagne, mais également pour interagir directement avec les électeurs. Plusieurs problématiques émergent.

Premièrement, les risques classiques liés à ces outils, tels que les biais discriminatoires ou les hallucinations, peuvent se manifester dans les interactions avec l'IA générative. Ce risque semble particulièrement important dans une situation où l'enjeu est de convaincre l'électeur, ce que le modèle sous-jacent au système peut être entraîné à faire grâce au RLHF par exemple (*reinforcement learning from human feedback*, une technique d'ajustement des modèles aux préférences des utilisateurs). Dans le cas d'un chatbot, seule la formulation des réponses pourra les rendre plus ou moins percutantes pour l'utilisateur, mais dans le cas d'une interaction orale, comme au téléphone, de nouvelles dimensions comme l'intonation ou le timbre de la voix pourront influencer la réception par l'interlocuteur. Ce cas d'usage n'appartient désormais plus à la science-fiction, puisque le candidat américain Peter Dixon y a eu recours sous les traits de « Jennifer », une IA développée par Civox. En amont des primaires démocrates aux Etats-Unis, cette « agente conversationnelle » a démarché les électeurs pour les informer sur la tenue de l'élection. Bien que l'IA informe immédiatement sur sa nature, les garanties apportées sur la qualité de ses réponses ne sont pas communiquées sur le site de l'entreprise Civox. On observe également des cas d'usages moins complexes techniquement, mais pouvant être tout aussi percutant en pratique, comme la diffusion de vidéos dont certains passages sont modifiés pour individualiser leur contenu. Cela a notamment été réalisé par un parti dans le cadre des élections indiennes auprès de bénévoles contactés sur Whatsapp.

Deuxièmement, les inégalités d'accès à de tels services pourraient défavoriser certains partis ou certains pays dans le cas d'une élection multinationale. D'une part, le coût du déploiement d'un dispositif de grande ampleur, comme l'agent conversationnel utilisé par Peter Dixon, pourrait être trop important pour permettre aux partis les moins dotés financièrement de s'en munir. Cette disparité aurait pour effet de réduire l'impact de ces petits partis, alors qu'aucune règle d'équité ne s'impose à la prospection politique en ce qui

concerne les communications directes auprès des électeurs comme cela est le cas pour la télévision et la radio comme le rappelle l'Arcom. D'autre part, les outils existants n'ont pas toujours les mêmes capacités dans toutes les langues en raison de la difficulté d'accéder à des jeux de données d'entraînement de volume suffisant pour certaines langues. En effet, des jeux volumineux et de qualité sont faciles à trouver pour des langues largement représentées dans le monde comme l'anglais ou le chinois alors que des langues européennes comme le letton ou le norvégien seront moins bien lotis, sans parler de certaines langues comme le romanche, bien qu'il s'agisse de l'une des quatre langues nationales suisses. Les extraits d'un recensement réalisé par le site Hugging Face ci-dessous montrent bien ces disparités dans le nombre de jeux de données existants pour chacune des langues (bien qu'il aurait été nécessaire de comparer le volume total de données pour chaque langue et la qualité des jeux pour réaliser une comparaison plus parlante).

Language	ISO code	Datasets	Models
English English	en	11,034	52,358
Chinese 中文	zh	1,238	4,583
French Français	fr	1,079	4,075
Spanish Español	es	966	3,287
...			
Norwegian Norsk	no	169	697
Swahili Kiswahili	sw	169	701
Latvian latviešu valoda	lv	161	484
Panjabi ਪੰਜਾਬੀ	pa	153	538
Indonesian	ind	136	30
...			
Javanese	jav	45	12
Ligurian	lij	44	48
Romansh rumantsch grischun	rm	43	50
Venetian	vec	43	66
South Azerbaijani	azb	43	46

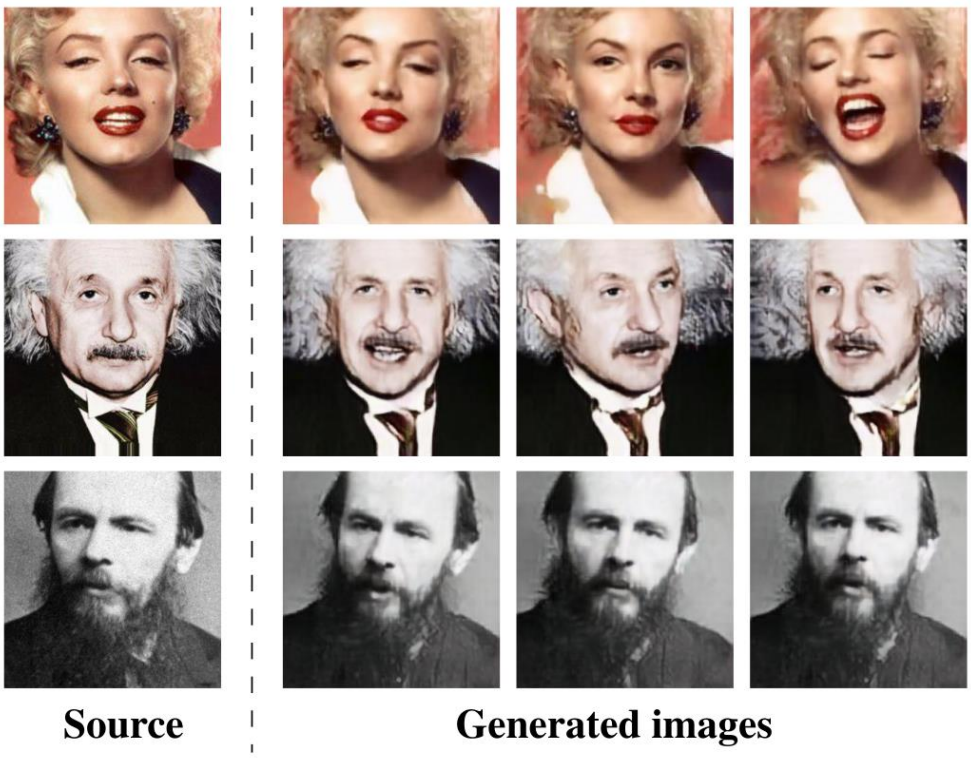
Recensement des jeux de données textuelles par langue pour l'anglais, le chinois, le français, l'espagnol, le norvégien, le swahili, le letton, le pandjabi, l'indonésien, le javanais, le ligure, le romanche, *le vénitien et l'azéri du sud (source : Hugging Face)*

Les disparités dans l'accès aux données influencent généralement la performance des modèles entraînés, au point où il pourrait être impossible de développer un agent conversationnel avec lequel l'interaction serait suffisamment fluide pour certaines langues. Le financement derrière les projets liés à des langues moins représentées à l'international peut également impacter la disponibilité des jeux de données et des modèles d'IA puisque le marché lié à leur commercialisation est *de facto* plus restreint.

Enfin, montré par une récente étude de l'organisation AI Forensics, ces contenus artificiels peuvent avoir pour objectif de dramatiser les récits portés par certains partis davantage que d'informer les électeurs sur leur programme. L'étude identifie 51 images artificielles postées sur les réseaux sociaux dont l'objectif est, selon les auteurs, d'amplifier des messages anti-UE et anti-migrants. Dans le cas étudié, la nature artificielle des contenus publiés n'était pas indiquée, ce qui questionne la responsabilité des partis, d'une part, mais également celle des plateformes les hébergeant et des services qui ont permis de les générer. Bien qu'il s'agisse de matériel de campagne, ces contenus semblent viser à un objectif supplémentaire exploré ci-dessous : l'influence du vote, notamment par la focalisation du débat public sur certains sujets sensationnalistes.

La génération de contenus pour influencer le vote et la participation

L'accessibilité maintenant acquise des outils de génération de contenus ouvre la porte à une amplification des campagnes de désinformation et d'influence. Leur utilisation par divers groupes (rivaux, activistes, acteurs étrangers) pourrait largement gagner en impact par la production de contenus en grand nombre, plus crédibles, spécifiques et diversifiés qu'auparavant. En effet, des acteurs étrangers n'ont dorénavant plus de difficulté à utiliser la langue du pays visé dans toutes ses subtilités, et notamment en utilisant les codes linguistiques des populations ciblées. Dans un article de The Verge, on apprend qu'OpenAI et Meta auraient tous deux identifié une campagne conduite par le ministère des affaires étrangères d'Israël afin d'inciter le Congrès américain à financer des actions militaires contre le Hamas. Le rapport de Meta de mai 2024 indique que les faux comptes alimentés par l'IA générative correspondaient à des profils assez spécifiques, comme des étudiants juifs vivant aux Etats-Unis, et publiaient des commentaires sur les pages d'organisations et de personnalités de premier plan. En ce qui concerne les images et vidéos, la qualité des contenus varie grandement : la vidéo de Donald Trump soutenant l'ancien premier ministre pakistanais Imran Khan possède encore plusieurs défauts, bien que la vidéo puisse tromper si elle est visionnée en faible qualité ou distraitement. Les vidéos de Léna Maréchal et d'Amandine Le Pen, des personnes fictives, supposément membres de la famille Maréchal-Le Pen, quant à elles, sont plus travaillées et utilisent notamment les codes des réseaux sociaux auxquelles elles sont destinées.



Exemple d'hypertrucages générés par l'IA à partir d'une unique photo (source : Zakharov et al., 2019)

L'intention derrière ces campagnes peut être multiple, et des stratégies et techniques différentes seront adoptées selon les cas. Dans le cadre d'élections, ces contenus peuvent viser à influencer le vote des électeurs, par divers moyens :

- en cherchant à les convertir à des opinions spécifiques, que ce soit par la confrontation fréquente à ces idées au moyen de comptes fictifs sur les réseaux sociaux, par la production de contenus convaincants comme des vidéos réalistes, ou par des hypertrucages convoyant un message via la bouche d'une personnalité reconnue,
- en dénigrant un candidat ou un parti par la propagation d'informations fausses,
- au contraire, en donnant de la visibilité à un candidat ou à un parti, notamment en augmentant l'engagement des utilisateurs de réseaux sur sa page (en générant des commentaires sur une page publique par exemple),
- en orientant le débat, par exemple en donnant de la visibilité à une opinion, à un parti, à des faits, ou encore en instaurant un climat particulier, afin de renforcer un sentiment (sécuritaire, libéral, ou de défiance envers les politiques par exemple) qui pourra orienter le vote des électeurs ou augmenter l'abstention.

Enfin, bien que l'on pense immédiatement à la propagation de vidéos ultraréalistes sur les réseaux sociaux lorsqu'on s'interroge sur l'utilisation de l'IA générative pour influencer le vote, il faut rappeler que la première utilité de ces outils est l'augmentation de la productivité. Comme souligné dans le rapport d'OpenAI sur les opérations d'influence, c'est également dans cet objectif que sont exploités ces techniques, augmentant l'efficacité et la portée des campagnes d'influence.

L'IA générative sous la plume du législateur

Concernant la création et le partage de contenu grâce à l'IA générative, plusieurs textes s'appliquent aux fournisseurs de ces services, aux plateformes pouvant héberger et relayer les contenus, ou encore pour les utilisateurs des services. Ces dispositions portent sur la transparence envers les destinataires, sur les mesures de marquage et de détection pour les plateformes, sur les mesures de réduction des hallucinations, ou encore sur les conditions de mise sur le marché de ces services lorsqu'ils sont considérés comme à haut risque.

Le Règlement relatif aux Services Numériques

Premièrement, le Règlement (UE 2022/2065) relatif à un marché unique des services numériques, ou RSN (fréquemment appelé DSA pour *digital services act* en anglais), comme les lignes directrices précédemment évoquées, vise à réguler les plateformes sur lesquelles les contenus artificiels pourraient être rencontrés. De nature contraignante, il prévoit par exemple que les fournisseurs de ces plateformes devraient agir contre les contenus illicites, notamment par un mécanisme de signalement, permettant aux utilisateurs d'identifier les publicités (dont les publicités politiques). De plus, les très grandes plateformes et moteurs de recherche (les VLOPs et VLOSEs), sont soumis à des obligations spécifiques. Ces « très grands » services sont désignés dans une liste publique par la Commission Européenne selon certains critères provenant du RSN. En mai 2024, il s'agissait pour les *very large online platforms*, ou VLOPs, notamment de Facebook, Instagram, Tiktok, X, Youtube, Linkedin, Wikipedia, et pour les *very large online search engines*, ou VLOSEs, de Google et Bing. Ces obligations spécifiques incluent notamment :

- la mise en place d'un système de gestion des risques visant à évaluer et à atténuer les risques systémiques liés à leur plateforme, dont l'impact sur le discours civique et les processus électoraux,
- l'accès aux données permettant de contrôler l'efficacité des mesures prises pour les chercheurs, ainsi que la tenue d'un registre des publicités promues sur leur plateforme.

La Commission européenne a également publié, en amont des élections européennes de juin 2024, des lignes directrices pour les très grandes plateformes et très grands moteurs de recherche sur l'atténuation des risques systémiques pour les processus électoraux afin de présenter les bonnes pratiques et recommander les mesures possibles aux très grandes plateformes et très grands moteurs de recherche en ligne. Les lignes directrices, qui anticipent les obligations du RIA et du règlement relatif à la transparence et au ciblage de la publicité politique, prévoient plusieurs catégories de mesures d'atténuation des risques, relatives :

- à l'éducation des utilisateurs et des destinataires des contenus (à laquelle la CNIL, en particulier via le Linc, contribue activement par ses publications), notamment sur le fonctionnement de l'IA générative et sur les utilisations abusives attendues,

- à la création de contenu, via l'utilisation de filigranes (technique liée aux diverses formes de tatouage numérique, auquel le Linc a dédié un article), aux métadonnées, ou encore à des méthodes cryptographiques permettant de prouver la provenance et l'authenticité. Ces mesures portent également sur la fiabilité et la traçabilité des sources, sur l'information des utilisateurs sur les erreurs potentielles, le test et la sécurité, en particulier par l'utilisation de filtres sur les entrées et sorties.
- à la diffusion de contenus, par la mise à disposition d'outils de marquage, l'obligation d'indiquer lorsque les contenus publiés sont artificiels, la détection des contenus manipulés par l'IA.

Le règlement sur l'intelligence artificielle

Le Règlement (UE 2024/1689) établissant des règles harmonisées sur l'intelligence artificielle, ou RIA, quant à lui, effectue une distinction selon le risque lié à chacun de ces systèmes. Certains systèmes au risque inacceptable sont ainsi interdits, comme les systèmes de notation sociale ou de manipulation des personnes visant à leur faire prendre une décision contre leur volonté. Sauf dans certains cas spéculatifs extrêmes, les systèmes évoqués plus haut ne rentrent *a priori* pas dans ces définitions. Viennent alors les systèmes à haut risque, soumis à certaines obligations préalables à leur mise sur le marché, parmi lesquelles on trouvera les systèmes destinés à être utilisés pour influencer le résultat d'une élection ou le comportement électoral de personnes dans l'exercice de leur vote lors d'élections. Néanmoins, sont exclus de cette catégorie les systèmes d'IA auxquels les personnes physiques ne sont pas directement exposées, tels que les outils utilisés pour organiser, optimiser ou structurer les campagnes politiques sous l'angle administratif ou logistique. Les IA génératives n'ayant généralement pas pour destination d'influencer le résultat d'une élection, beaucoup n'entreront pas dans cette définition, alors que leur utilisation est susceptible d'avoir ces mêmes effets. Enfin, des dispositions portent spécifiquement sur l'IA générative, parmi lesquelles :

- des obligations de transparence pour les fournisseurs et déployeurs de service, visant à assurer que les personnes ayant à interagir avec les contenus artificiels en soient informés, notamment par leur marquage. Cette obligation rejoint l'article 15 de la loi visant à sécuriser et à réguler l'espace numérique (SREN) qui interdit la publication d'hypertrucages sans le consentement de la personne visée s'il n'est pas clairement indiqué que le contenu est artificiel,
- des obligations pour les fournisseurs de modèles à usage général, reposant principalement sur la documentation, et dont les modèles publiés en source ouverte sont exemptés,
- des obligations pour les systèmes à risque systémique (selon certains critères donnés par le RIA), incluant les risques pour les processus électoraux et la dissémination de contenus faux, illégaux ou discriminatoires. Ces obligations incluent des mesures pour identifier et atténuer ces risques comme l'évaluation du modèle, ou l'analyse, le suivi et la documentation des risques et incidents.

Au stade actuel, plusieurs interrogations persistent, notamment sur la faisabilité du marquage des contenus, le tatouage numérique manquant encore de robustesse comme décrit dans un article dédié, ou encore sur l'identification des systèmes à risque systémique. Ces derniers sont désignés par le RIA par des critères techniques novateurs liés à leurs capacités et à leur portée, qu'il conviendra d'évaluer lors de l'entrée en vigueur du règlement.

Conclusion

Un cadre règlementaire existe donc déjà pour limiter les risques liés à l'utilisation de l'IA générative pour influencer le résultat d'élections, bien que la majorité de ces dispositions aient encore à entrer en vigueur. De plus, lorsqu'il s'agit d'obligations de moyens et non de résultats, les limitations de la technique pourraient rendre inefficaces ces mesures (comme dans le cas de la détection des contenus artificiels). En pratique, l'IA générative n'est qu'un seul des deux piliers sur lesquels reposent les stratégies de désinformation. Les systèmes de recommandation en constituent le second.

Des annonceurs aux utilisateurs : un cheminement algorithmique [2/3]

Les systèmes de recommandations développés et utilisés par les réseaux sociaux et les moteurs de recherche visent en premier lieu à fournir du contenu et des publicités adaptés à l'utilisateur tout en servant les intérêts économiques des fournisseurs de service et des annonceurs publicitaires. Comme décrit dans la suite, le système de recommandation de contenu (ce qui alimente un flux sur un réseau social ou les réponses sur un moteur de recherche) constitue un premier composant technique, dont la compréhension permet de mieux cerner les risques liés aux systèmes de ciblage publicitaire.

Des systèmes de recommandation toujours plus intelligents

L'objectif sous-jacent à la recommandation de contenu est l'optimisation de « l'engagement » des utilisateurs, qu'il s'agisse du temps passé sur la plateforme, ou du nombre de mentions, de commentaires, de partage d'un contenu ou du nombre d''achats effectués pour les plateformes de vente. Bien que leurs concepteurs soient rarement transparents sur la transposition technique de cet objectif dans le fonctionnement de ces systèmes, certains recensements, comme Campana & Delmastro, 2023, en proposent une classification également étudiée dans le cahier IP « Les données, muses et frontières de la création » :

- **le filtrage collaboratif**, où les interactions des utilisateurs avec les contenus sont utilisées pour leur recommander de nouveaux contenus. Ces algorithmes peuvent être :
    - basés sur la mémoire : c'est-à-dire que l'historique des utilisateurs ayant interagi positivement avec un contenu sera utilisé pour effectuer une recommandation à un utilisateur dont l'historique est similaire,
    - basés sur un modèle : c'est-à-dire qu'une phase de modélisation a lieu sur la base des interactions des utilisateurs, permettant de recommander de nouveaux contenus. Ce type d'algorithmes repose généralement sur la création d'une représentation mathématique des contenus (via un espace latent),
- **le filtrage basé sur le contenu**, où une description explicite du contenu est utilisée afin de le recommander à un utilisateur ayant indiqué son intérêt pour cette catégorie (explicitement, via un profil d'utilisateur, ou implicitement, via son historique).

A ces deux grandes catégories doivent être ajoutées certaines techniques pouvant exploiter le contexte (comme l'heure du jour, la géolocalisation, etc.), les choix de l'utilisateur (sur Facebook par exemple, c'est le rôle des boutons « show more » et « show less » associés aux publications), ou encore l'apprentissage actif (par renforcement par exemple, comme étudié par des chercheurs de Google). Clustering, réseaux de neurones, réseaux en graphes, etc. les

algorithmes utilisés varient et peuvent se composer afin d'utiliser plusieurs des approches précédentes de manière hybride.

Ces algorithmes sont encore très utilisés aujourd'hui pour se connecter avec d'autres personnes et pour accéder à l'information en particulier devant la quasi infinité de contenus disponibles et dont la qualité est grandement variable. Le EU Internet Forum (un consortium supervisé par des députés européens et regroupant des membres privés, publics et de la société civile) s'est attaché à évaluer l'impact de l'amplification algorithmique des contenus Terroristes, Violents et Extrémistes (TVE) ainsi que des contenus limites (*borderline* en anglais). Les conclusions de cette étude ont été publiées fin 2023. Les contenus *limites* sont caractérisés par leur capacité à conduire à la radicalisation sans pour autant être illégaux (via la désinformation, les thèses conspirationnistes, ou en cherchant à manipuler les utilisateurs). L'étude conclue que l'interaction avec des contenus TVE ou *limites* contribue à leur amplification via les algorithmes de recommandation. Elle observe notamment :

- une amplification de 18% en un jour des contenus TVE
- une amplification de 65% en un jour pour les contenus limites,
- que 90% des contenus signalés aux plateformes étaient encore en ligne 8 semaines après le signalement.

Ces résultats peuvent varier selon les langues, les plateformes, et les catégories de contenus (terroristes, violent d'extrême droite ou gauche, etc.). Les contenus limites en français en particulier seraient particulièrement amplifiés, 8 fois plus de contenus étant recommandé 24h après des recherches visant spécifiquement ces contenus. De plus, lorsque les contenus limites sont effectivement retirés des plateforme (ce qui ne concerne qu'environ 10% des cas), cela peut avoir lieu plusieurs semaines après le signalement, comme le montre le schéma ci-dessous tiré de l'étude.

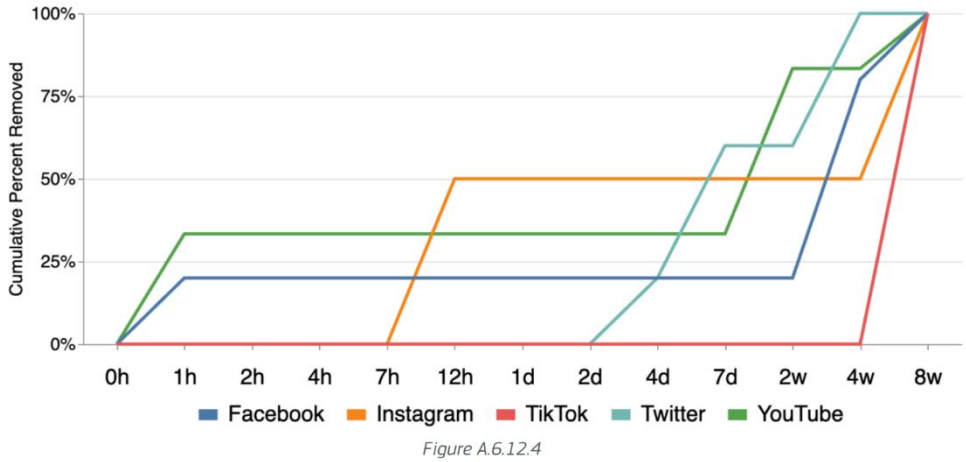


Figure A.6.12.4

This chart shows how long it took for each platform to remove the total amount of Borderline TVE content. The line graphs show the cumulative percentage, ending in 100% at the top right corner. This chart does not take into account any Borderline TVE content that wasn't removed.

Même si cette étude possède ses limites (notamment dues à la difficulté d'accéder à des données à grande échelle tel que souligné dans le document), elle démontre assez clairement que les réseaux sociaux permettent l'amplification de contenus pouvant

influencer un vote, que ce soit par la désinformation, en conduisant les utilisateurs à la radicalisation ou en les enfermant dans des bulles de filtre.

Les constats précédents semblent pointer l'insuffisance des mesures prises par les plateformes pour contrer les effets de l'amplification algorithmique des contenus préjudiciables, comme le signalement des contenus par les utilisateurs. D'autres pistes, reposant parfois elle-même sur l'intelligence artificielle, sont explorées comme nous le verrons dans le troisième article de cette série. Le droit pourrait également ouvrir la porte à davantage de transparence et de contrôle par les utilisateurs sur les contenus qui leur sont recommandés.

Quelles recommandations pour les systèmes de recommandation ?

Les lignes directrices pour les très grandes plateformes et moteurs de recherche prévoient des mesures d'atténuation des risques incluant :

- La possibilité, via une option à la main de l'utilisateur de choisir et de maîtriser son flux d'information,
- La réduction de la désinformation, par la limitation de l'amplification de ces contenus,
- La modification des algorithmes lorsqu'ils font peser des risques sur les processus électoraux,
- La transparence sur la conception et le fonctionnement des systèmes de recommandation, et la coopération pour leur audit.

Ces recommandations rejoignent des obligations prévues par le *Règlement sur les services numériques* (RSN) qui impose aux fournisseurs de service d'indiquer les principaux paramètres utilisés dans les systèmes de recommandation dans les conditions d'utilisation, ainsi que les options dont disposent les utilisateurs pour les modifier ou les influencer. Ce texte prévoit également qu'au moins une option qui ne reposerait pas sur du profilage leur soit proposée. Enfin, le système de gestion des risques (dont l'impact sur les processus électoraux) à mettre en place par les plateformes doit inclure une évaluation de l'influence des systèmes de recommandation.

Comme évoqué dans l'article précédent, le règlement sur l'IA (RIA) prévoit des obligations pour les systèmes selon leur niveau de risque. Néanmoins, les algorithmes de recommandation ne semblent pas entrer dans les catégories à haut risque. Des clarifications à ce sujet pourraient être apportées par la Commission Européenne.

Enfin, le RGPD prévoit que les risques de discrimination soient pris en compte dans l'analyse d'impact sur la protection des données (AIPD) liée au traitement de données personnelles mis en œuvre par un système de recommandation. Ces risques doivent ensuite être réduits à un niveau acceptable, sans quoi l'AIPD doit être envoyée à l'autorité de protection des données. Puisque l'opinion politique est l'un des critères sur lesquels une discrimination peut être caractérisée selon le Défenseur des Droits, faire entrer une personne dans une bulle de filtre sur la base de ses interactions passée avec des contenus à caractère politique pourrait être considéré comme une discrimination. De plus, le RGPD prévoit des dispositions

spécifiques au profilage prévues à son article 22, lorsque celui-ci possède un impact significatif sur les personnes. Cet article impose des conditions pour qu'une prise de décision puisse être fondée exclusivement sur le résultat donné par un système automatisé, comme une IA. Toutefois, les notions correspondantes à la prise de décision purement automatisée, ou à ce qui constitue un impact significatif restent à interpréter puisque la jurisprudence portant sur cet article est encore mince : à la connaissance de l'auteur, aucun jugement n'a considéré qu'un algorithme de recommandation remplissait ce critère jusqu'à aujourd'hui. Le G29 (le groupe qui réunissait les autorités de protection des données européennes jusqu'en 2018, aujourd'hui devenu l'EDPB) ne l'exclue pas en théorie dans ses lignes directrices sur le sujet, selon les caractéristiques particulières de la situation, y compris en ce qui concerne :

- le caractère intrusif du processus de profilage, notamment le suivi des personnes sur différents sites web, appareils et services ;
- les attentes et les souhaits des personnes concernées ;
- la façon dont l'annonce est diffusée ; ou
- le recours aux vulnérabilités connues des personnes concernées visées.

Les dispositions prises en amont visent ainsi principalement trois objectifs : la reprise de contrôle par les utilisateurs, la transparence, ainsi que l'analyse de risque. Les techniques permettant de répondre à ces enjeux sont de divers ordres :

- l'inclusion des choix de l'utilisateur dans le système de recommandation, en tenant compte de ses retours sur les contenus proposés (pour tenir compte de leurs goûts, et des signalements de contenus inappropriés), de certains intérêts qu'il indiquerait, ou encore en lui proposant de choisir entre plusieurs algorithmes de recommandations,
- la prise en compte de critères supplémentaires dans l'algorithme tels que la diversité des contenus, ou leur sérendipité (c'est-à-dire à quel point un nouveau contenu pourrait surprendre positivement l'utilisateur),
- l'identification des contenus illégaux, préjudiciables ou limites, ainsi que des tentatives d'influence (voir l'article suivant concernant ces techniques et leurs limitations),
- la transparence, notamment sur le fonctionnement des algorithmes, ainsi que sur les contenus recommandés afin de permettre l'évaluation par des tiers.

Ces nouvelles mesures ne peuvent conduire que dans la bonne direction, mais les premières évaluations du système de gestion des risques prévu par le RSN incitent à la prudence. Dans une publication étudiant l'application de ce système aux campagnes de désinformation russes sur les plateformes, la Commission Européenne dresse un bilan très négatif des mesures prises. En effet, malgré le blocage des comptes de représentants politiques russes et la modération des contenus partagés sur les réseaux, la désinformation russe liée à la guerre en Ukraine semble toujours atteindre un public en Europe en partie à cause des systèmes de recommandation. Grâce à un indicateur appelé « Non-Follower Engagement », il a pu être montré que si la portée des publications des influenceurs et médias russes bloqués sur les réseaux sociaux avait diminué, celle des publication des employés du Kremlin

et des représentants officiels (comme les ambassades) quant à elle, avait augmenté de manière à compenser la première baisse.

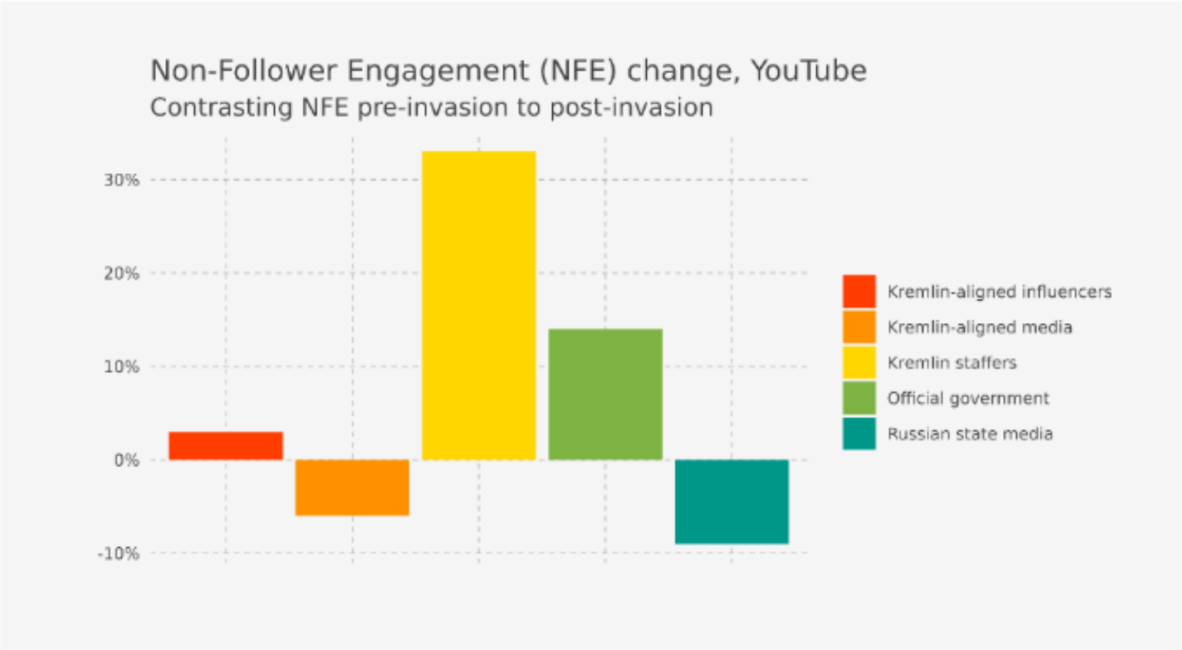


Figure 26: NFE pre-invasion (1 December 2021 – 20 February 2022) contrasted to NFE post-invasion (1 April 2022 – 30 November 2022) on YouTube.

Source : Commission Européenne, [Application of the risk management framework to Russian disinformation campaigns]

D'après l'étude, ceci pourrait être expliqué d'une part par un déport des utilisateurs vers les comptes toujours autorisés sur les plateformes, leur donnant davantage de visibilité dans les recommandations. D'autre part, les utilisateurs pourraient être conduits vers ces comptes par d'autres biais, par exemple par les moteurs de recherche, ou par des boucles de diffusion plus restreintes comme des groupes de discussion ou forums. Il apparaît ainsi que des mécanismes sous-jacents, tels que ces boucles de rétroaction, existent et échappent encore à la modération mise en œuvre par les plateformes, en partie à cause d'un manque de contrôle sur les algorithmes de recommandation. Ce constat semble d'autant plus préoccupant en raison de la maîtrise de ces mécanismes acquise par certains acteurs. En effet, une étude récente du chercheur David Chavalarias réunit une liste d'indices conduisant à penser que les acteurs russes en particulier utilisent des techniques comme l'astroturfing (amplification artificielle d'une idée par la création d'une foule factice la propageant) ou les publicités ciblées dans le cadre d'une stratégie d'ingérence élaborée sur le long terme.

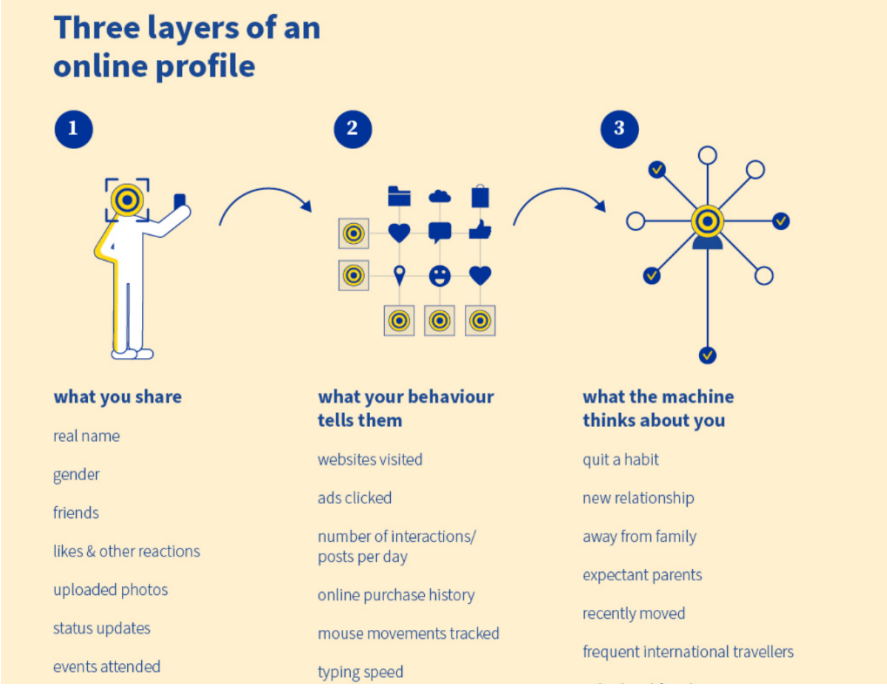
Recommandation et ciblage publicitaire : du pareil au même ?

D'après une analyse conduite sur le cas particulier de Facebook par Chouaki et al., 2022 , les annonceurs ont le choix entre deux stratégies pour cibler leur audience :

- le ciblage à la main de l'annonceur : où il choisit lui-même les caractéristiques des groupes de personnes qu'il souhaite toucher. Ces caractéristiques peuvent reposer

sur des informations démographiques déclarées par l'utilisateur (comme son âge, son sexe, son lieu de résidence, etc.), sur des « segments » inférés par la plateforme à partir des interactions des utilisateurs sur ses produits et cherchant à décrire ses intérêts (comme un intérêt pour le cinéma, pour un sport ou un genre de musique – les segments liés à un compte Facebook peuvent être trouvés dans la rubrique « ad topics »),

- le ciblage algorithmique : les techniques utilisées sont similaires aux algorithmes des systèmes de recommandation décrits plus haut.



Infographie décrivant le fonctionnement du profilage en ligne (source : Conseil de l'UE)

Ces deux stratégies de ciblage semblent assez différentes, la première laissant une marge de manœuvre importante aux choix de l'annonceur, et la seconde reposant sur un algorithme (avec des limitations similaires à celles des systèmes de recommandation). Cependant, comme indiqué sur une page d'aide de Meta dédiée aux annonceurs, lorsqu'un ciblage « manuel » est choisi, le ciblage peut toujours avoir lieu sur la base du résultat algorithmique lorsque cela peut « améliorer les performances ». Ce ciblage est mis en œuvre via un système d'enchères tenant compte de trois critères : le montant de l'enchère, le taux d'interactions estimé et la qualité de la publicité.

Une étude conduite en 2019 (donc sur la base d'un dispositif de ciblage publicitaire qui a pu évoluer depuis) par Ali et al., 2019 a démontré qu'en ce qui concerne le ciblage automatique des publicités politiques aux Etats-Unis, le système utilisé par Facebook avait pour effet :

- de délivrer préférentiellement les publicités politiques aux utilisateurs dont les opinions sont proches de celles portées par la publicité,
- de limiter la possibilité (ou d'augmenter le coût) pour les annonceurs d'atteindre une audience qui ne partagerait pas ses opinions politiques.

Ces constats auraient pour effet d'amplifier les bulles de filtre dans lesquelles les utilisateurs se retrouvent bloqués, et ainsi d'amplifier la polarisation politique. Par ailleurs, le coût lié au ciblage publicitaire d'une part, et le surcoût lié à l'acquisition d'une audience d'un autre bord politique d'autre part, pourraient avoir pour double effet de diminuer la visibilité des partis les moins bien financés, et d'augmenter la capacité des plus gros partis à convertir des électeurs. Ces effets ont sans aucun doute des effets sur les opinions politiques majoritaires, qu'il reste à évaluer.

Une infographie du Conseil de l'UE indique que 43 millions d'euros avaient été dépensés pour la publicité politique en ligne au cours des 20 semaines précédant l'élection européenne de 2019. Ce montant est toutefois très faible en comparaison des dépenses annuelles au niveau mondial : 522.5 milliards de dollars en 2021, et il devrait encore augmenter pour atteindre 836 milliards de dollars en 2026. En Europe, la France est l'un des marchés les moins importants en dépenses – 321k€ entre mars 2019 et janvier 2022 – en comparaison d'autres pays comme l'Allemagne ou la Hongrie – respectivement 5,35M€ et 2.96M€ sur la même période. Cela peut s'expliquer d'une part par l'interdiction prévue par le code électoral d'utiliser tout procédé de publicité commerciale à des fins de propagande électorale 6 mois avant un scrutin. Bien que la présence ou non d'échéances électorales sur la période dans les pays comparés, ainsi que la crise du Covid aient nécessairement influencé ces montants, la différence d'ordre de grandeur est tout de même parlante. Puisque les réseaux sociaux sont principalement financés par les espaces de publicité (politique ou non) qu'ils monétisent, ces sommes sont représentatives d'importants enjeux pour les fournisseurs. Elles démontrent également l'importance de l'espace numérique pour la communication politique des partis.

Réglementer les publicités politiques : un enjeu de transparence

Etant ainsi admis que ce vecteur est un moyen d'information important pour les citoyens, les textes portant sur la publicité politique ne visent pas à l'interdire, mais à lui fixer un cadre.

Le règlement sur la publicité à caractère politique, tout d'abord, introduit des critères permettant de déterminer les contenus relevant de la publicité politique (tenant compte du contenu, du parraineur, du langage, ou encore du contexte de publication). Ce texte interdit le traitement de données personnelles et le ciblage lorsque les données n'ont pas été collectées auprès de la personne concernée (notamment par le moissonnage sur le Web), lorsque la personne n'a pas donné son consentement, lorsque le profilage est basé sur des données sensibles (visées à l'article 9 du RGPD), ou encore lorsque la personne concernée ne sera pas en âge de voter avant encore un an. Il prévoit également plusieurs mesures de transparence :

- sur les publicités politiques, que les éditeurs devront obligatoirement marquer, et indiquer aux utilisateurs notamment l'identité du parraineur, et l'utilisation éventuelle de techniques de ciblage ;
- la tenue par la Commission d'un répertoire des publicités politiques accessible au public ;

- sur les techniques de ciblage et de diffusion des publicités politiques, les éditeurs devront adopter et publier des règles internes sur leur utilisation, transmettre des informations sur leur logique et leurs principaux paramètres. Il devrait notamment être précisé si un système d'intelligence artificielle est utilisé, les groupes de destinataires ciblés, ou encore les paramètres utilisés pour le ciblage (dont les catégories de données personnelles utilisées).

De plus, les éditeurs devront préparer une évaluation interne annuelle des risques que l'utilisation des techniques de ciblage ou de diffusion représentent pour les libertés et les droits fondamentaux, et en publier les résultats.

Ces dispositions seront applicables à partir de l'automne 2025 dans leur majorité. Elles viendront ainsi compléter les mesures prévues par le RSN. Ces dernières portent plus généralement sur tous types de publicités et prévoient que les fournisseurs de très grands moteurs de recherche et plateformes en ligne tiennent et publient un registre des publicités. Il devrait notamment y être indiqué l'identité des personnes ayant financé et demandé la diffusion (lorsqu'elles sont distinctes), si certains groupes de destinataires étaient ciblés lors de la diffusion, et les principaux paramètres utilisés à cette fin (comme les paramètres utilisés pour exclure certains groupes particuliers). Le texte prévoit également des mesures d'audit par les Etats membres (via des coordinateurs nationaux : l'Arcom en France) et par des chercheurs agréés par un accès aux données nécessaires. Les coordinateurs nationaux pourront également demander la fourniture d'informations sur le fonctionnement et les tests effectués sur les systèmes algorithmiques.

Toutefois, en dépit de ces avancées règlementaires importantes sur l'encadrement des publicités politiques, une difficulté importante persiste comme souligné par Oana Goga dans une interview donnée au Linc en 2021 : un nombre important de publicités politiques circuleraient sur Facebook sans être identifiées comme telles. Ainsi, la présence de ces contenus sur les réseaux et plateformes pourrait entraîner les conséquences discutées plus haut en termes de discriminations, de bulles de filtre et de désinformation, sans que les mesures sur la transparence, le ciblage et la qualité de l'information ne soient applicables. Des techniques et mesures existent pourtant pour identifier ces contenus comme décrit dans l'article suivant.

L'autocontrôle, un nouvel apprentissage pour l'IA [3/3]

Face à la prolifération de contenus liés à la désinformation, qualifiés de TVE (terroriste, violents ou extrêmes, voir l'article précédent), ou visant à influencer les électeurs (hypertrucages, publicités politiques non qualifiées comme telles), ainsi qu'aux risques des systèmes algorithmiques utilisés pour la recommandation ou le ciblage et pouvant être exploités par des acteurs étrangers (au moyen de comptes automatisés appelés bots notamment), des mesures de contrôle semblent plus que nécessaires. Certaines sont d'ores et déjà mises en œuvre par les plateformes elles-mêmes ou par les Etats. Les tiers, comme la société civile et les chercheurs, jouent également un rôle crucial dans la régulation en pointant les risques ou en identifiant les manquements des plateformes à leurs obligations.

Les autorités de contrôle publiques

Un tour d'horizon réalisé par l'organisation IFES montre que les travaux visant à construire la résilience contre les opérations d'influence en amont des élections européennes de 2024 sont très nombreux. Parmi eux, certains touchent à l'amélioration de la qualité de l'information dispensée par les médias ou encore à l'éducation des citoyens, mais seuls la France et la Suède semblent avoir créé des agences dotées de compétences techniques : Viginum et la Psychological Defense Agency. La commission européenne quant à elle repose principalement sur le financement de projets tels que le Consortium EDMO (European Digital Media Observatory) pour lutter en pratique contre la désinformation. Les services de la Commission s'attachent ainsi davantage à surveiller l'application des textes, avec l'ECAT (European Center for Algorithmic Transparency) en ce qui concerne le RSA, ou le bureau de l'IA pour ce qui est du Règlement IA. Ces tâches sont parfois déléguées à des autorités nationales, comme dans le cas du RGPD, ou de certaines dispositions du RSN.

En France, Viginum a pour mission de détecter et caractériser les ingérences étrangères via l'étude de phénomènes inauthentiques sur les plateformes (comme les comptes suspects, les contenus malveillants, les comportements anormaux, aberrants ou coordonnés). Cette unité du Secrétariat Général de la Défense et de la Sécurité Nationale, qui compte plusieurs *data scientists*, conduit des recherches sur les techniques de détection des ingérences étrangères. Elle a récemment publié un article proposant un cadre pour l'identification et la catégorisation des bots sur X grâce à l'analyse des données, sur la base de leurs caractéristiques et de leur comportement. Les travaux d'un autre organisme public, le PEReN, ou Pôle d'Expertise de la Régulation Numérique viennent appuyer les missions des autorités de surveillance et de régulation comme Viginum ou l'Arcom en leur fournissant des outils permettant par exemple de détecter les contenus artificiels, d'analyser les rapports

sur la modération de contenu fournis par les plateformes, ou encore d'analyser les bulles de filtres créées par les algorithmes de recommandation.

Bien que ces travaux portent leurs fruits, la caractérisation par Viginum du réseau pro-russe « Portal Kombat » en est une preuve, la mise en œuvre de ces techniques par les autorités publiques n'est pas sans risque pour la vie privée des usagers des plateformes. Comme l'a rappelé la CNIL dans son avis sur la création de Viginum, l'existence de cette surveillance peut avoir des conséquences pour les utilisateurs et notamment modifier leurs comportements en ligne. Par ailleurs, si ces services sont jugés nécessaires et proportionnés dans une société démocratique, le risque qu'ils soient détournés de leur finalité initiale existe, soit par un acteur malveillant (un employé peu scrupuleux qui parviendrait à contourner les mesures de sécurité prévues par exemple), soit en raison d'une évolution vers une société de plus en plus sécuritaire. Ce dernier risque est d'ores et déjà documenté et déploré par certains mouvements. Comme l'indiquait une experte indépendante des Nations Unies : « Les justifications exceptionnelles de l'utilisation des technologies de surveillance dans le cadre de la lutte antiterroriste "allégée" des droits de l'homme se transforment souvent en une utilisation régulière banale ».

Le contrôle par les plateformes

Additionnellement aux dispositifs prévus par les autorités publiques, les plateformes ont depuis plusieurs années cherché à développer leurs propres outils de contrôle. Ces outils peuvent répondre à des obligations légales, ou servir d'autres objectifs (comme l'amélioration du service pour les utilisateurs, ou le retrait de certains contenus dont la présence sur une plateforme pourrait avoir des conséquences en termes d'image pour son fournisseur).

Ces obligations légales sont pour le moment prévues principalement par *le règlement sur les services numériques* (RSN). Bien que ce dernier exclue l'obligation pour les fournisseurs de détecter activement les contenus illégaux, il les oblige à répondre aux injonctions des autorités concernant certains contenus illicites. De plus, ils doivent prévoir plusieurs mesures :

- la mise en place de mécanismes publics de signalement des contenus illicites,
- la suspension temporaire du service fourni aux utilisateurs publiant des contenus illicites,
- pour les très grandes plateformes et moteurs de recherche en ligne uniquement, l'analyse des risques systémiques dont la diffusion de contenus illicites, et des effets de leurs systèmes de modération des contenus et la mise en place de mesures d'atténuation raisonnables, proportionnées et efficaces.

Ainsi, les plateformes se dotent d'outils de détection de contenus inauthentiques, de manière individuelle, à la manière de Tiktok, ou coordonnée au moyen d'alliances comme le Global Internet Forum to Counter Terrorism (GIFCT). Le GIFCT dont la raison d'être première est la lutte contre le terrorisme sur les plateformes, s'attache également à prévenir la diffusion de contenus limites (*borderline*, voir l'article précédent). Dans un rapport de juin

2023, il liste les outils développés dans cet objectif, comme une base de hash (signatures d'images limites connues), un outil de comparaison d'une image avec les hash, des outils utilisant l'IA pour détecter les contenus terroristes, ainsi que des solutions pour aider à la modération humaine et au retrait des contenus. D'autres acteurs développent des outils dédiés spécifiquement à la détection d'hypertrucages, comme FakeCatcher d'Intel ou APATE (A Prototype Assessment Toolbox for Forensic Experts) d'Idemia. D'après les concepteurs de ces outils, leur exactitude dépasserait les 95% ce qui semble conforme aux ordres de grandeurs obtenus par Yan et al., 2023. Ces résultats devraient toutefois être confirmés sur la durée d'une part, en raison de l'évolution des techniques de génération, et sur différentes catégories d'hypertrucages d'autre part (pour le remplacement du visage uniquement, la synchronisation des lèvres, etc.). Ces résultats théoriques ne tiennent pas non plus compte des stratégies mises en œuvre par les acteurs cherchant à propager les contenus illicites, comme l'altération de certains mots clés (« U. kr. ai. n. e. » remplace « Ukraine » par exemple) identifiée dans le rapport de Meta de mai 2024 sur l'interception de menaces.

Le rôle des tiers

C'est enfin par les tierces parties, tels que les chercheurs académiques et les représentants de la société civile, que peut s'effectuer le contrôle des mesures prises par les plateformes en ligne.

La difficulté de ces acteurs à accéder aux données des plateformes, devrait être résolue par le RSN qui prévoit un accès pour les chercheurs agréés. Néanmoins, le plus souvent, les résultats mis à disposition sont partiels ou manquent de représentativité (bien qu'ils suffisent souvent à identifier un risque réel). Les moyens mis à disposition par les plateformes varient d'un service à l'autre et n'offrent généralement pas suffisamment de stabilité et de transparence comme le relève NiemanLab notamment. En réaction, les chercheurs mettent en œuvre diverses stratégies pour accéder aux données, dont certaines sont recensées dans une publication en cours de validation. L'auteur, Jimi Adams, relève plusieurs catégories : les expérimentations, les questionnaires, l'observation, et l'analyse des traces comportementales, chacune ayant ses avantages et limitations selon les objectifs de l'étude.

Malgré ces obstacles, des outils sont proposés par ces acteurs comme ContentCheck, un logiciel de vérification de l'information développé par un consortium de chercheurs en lien avec le journal Le Monde. Ce dispositif exploite l'intelligence artificielle afin de représenter les sources d'informations selon des graphes et d'en tirer les informations les plus pertinentes.

Enfin, c'est sur les utilisateurs des plateformes que repose souvent la modération des contenus par les signalements qu'ils peuvent effectuer sur les plateformes. Toutefois, ces signalements doivent être analysés par les plateformes ou leurs prestataires. Plusieurs organisations et chercheurs ont dénoncé les conditions de travail des employés de ces prestataires observées notamment via l'amplification de dynamiques discriminatoires et la faible rémunération des travailleurs, comme souligné par le UCD Center for Digital Policy, ou à cause de l'impact psychologique qu'entraîne l'exposition aux contenus, comme étudié par

des chercheurs de l'université londonienne de Middlesex. Par ailleurs, l'étude du EU Internet Forum mentionnée dans le deuxième article de cette série a montré que ces signalements ne conduisent que dans une minorité des cas au retrait des contenus, et après une certaine période. Les rapports prévus par le RSN que les fournisseurs de services devront publier une fois par an devraient apporter un éclairage sur ces pratiques, en indiquant par exemple le nombre de signalement des utilisateurs, le nombre de contenus supprimés, ou encore le taux d'erreur des systèmes automatisés de modération.

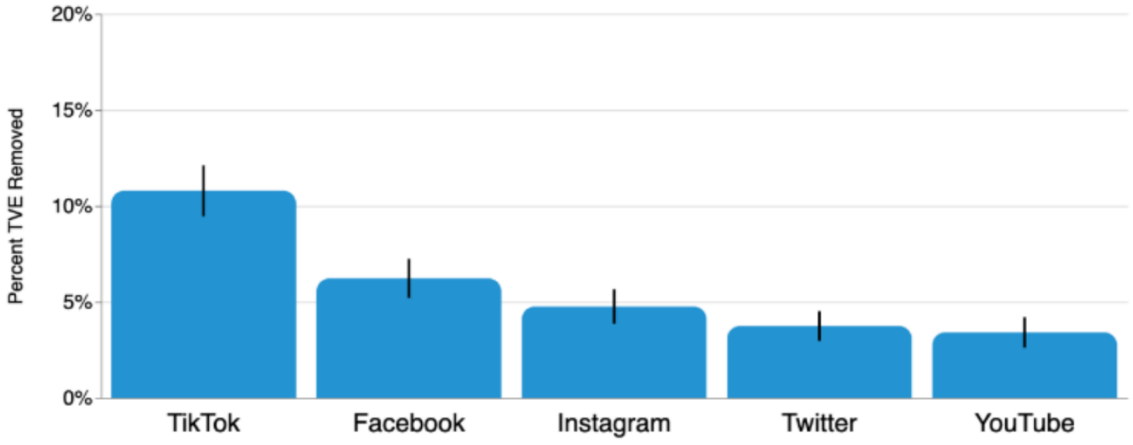


Figure A.6.8.1

This graph shows the percentage of TVE content that was removed on a given platform. The black lines in the bars represent 90% confidence intervals.

Source : EU Internet Forum, Study on the role and effects of the use of algorithmic amplification to spread terrorist, violent extremist and borderline content : final report

Le droit sera-t-il un outil suffisant pour protéger la démocratie ?

En conclusion, les utilisations de l'IA possèdent visiblement une influence sur les élections, tant en ce qui concerne la production de contenus visant à modifier les comportements des électeurs, la modération de ces contenus, ou la recommandation et le ciblage des contenus et publicités. Le cadre règlementaire, au travers des divers textes évoqués, cherche à s'adapter à ces évolutions. Les évolutions en faveur de l'éducation à ces risques, de la transparence sur les algorithmes et le contrôle des contenus devrait entraîner des effets positifs sur l'influence que les plateformes peuvent avoir sur l'opinion des électeurs, bien que ce contrôle ne soit pas sans effet sur la vie privée des utilisateurs. Toutefois, il sera nécessaire d'observer les élections à venir afin de vérifier si ces changements permettront d'éviter des écueils en situation réelle. Les périodes électorales étant souvent liées à une médiatisation importante parfois hâtive, les mesures prises pourraient ne pas avoir l'effet escompté dans un délai suffisamment court pour empêcher des conséquences regrettables. Ces limitations ont été démontrées par la diffusion sur Facebook de l'hypertrucage audio d'un candidat slovaque deux jours avant les élections législatives du pays. Bien qu'aucun lien de cause à effet ne puisse être établi, le candidat visé a terminé deuxième alors que certains sondages prévoyaient une victoire du parti. Les nouvelles mesures permettront-elles de

détecter, de qualifier de désinformation, et de retirer des plateformes à temps les hypertrucages d’une même nature lors des prochaines élections ?