Arcom

Le régulateur de la communication audiovisuelle et numérique

Lutte contre la manipulation de l'information sur les plateformes en ligne

Bilan annuel des moyens et mesures mis en œuvre pour protéger l'intégrité des services et lutter contre les techniques de manipulation

Mars 2026

© Arcom

Sommaire

Résumé exécutif

Initialement, les acteurs agissant sur le territoire européen – plateformes, moteurs de recherche, médias, société civile, chercheurs et autorités publiques – se sont fédérés dans une démarche d'adhésion volontaire pour mettre en place un code de bonnes pratiques, renforcé en 2022, contre la désinformation. Le règlement sur les services numériques (RSN, *Digital services Act* ou DSA en anglais), entré en vigueur en 2024 a permis d'approfondir cette démarche en rendant opposables les engagements volontaires pris par les plateformes et moteurs de recherche signataires, dans le cadre d'un nouveau code de conduite contre la désinformation, intégré au RSN en 2025. L'adhésion volontaire à ce code fait partie des mesures d'atténuation des risques systémiques prévues par le règlement, pour prévenir les phénomènes de manipulation de l'information. Le code et le règlement imposent concomitamment que les mesures mises en œuvre pour protéger l'intégrité des plateformes et lutter contre les techniques de manipulation soient respectueuses de l'impératif de protection de la liberté d'expression et d'information des utilisateurs en ligne.

À ce titre, les grandes plateformes ne focalisent pas leur action sur les contenus eux-mêmes mais sur les comportements et les techniques de certains acteurs qui relèvent de la manipulation de l'information. Chaque plateforme adapte son action aux caractéristiques de son service, de son interface, de ses algorithmes et des menaces qui peuvent en découler. Il n'est donc pas aisé d'avoir une vision globale de l'ensemble des actions menées. Le présent rapport s'attache à pallier cette difficulté en ce qui concerne les principales plateformes : Facebook et Instagram (Meta), Google Search, YouTube, Microsoft Bing, LinkedIn, X, TikTok, Snapchat et Wikipédia.

Dans un premier temps, le rapport présente les tactiques, techniques et procédures (*TTPs*) de manipulation de l'information en ligne telles qu'elles résultent des travaux réalisés dans le cadre du code de conduite. Dans un deuxième temps, et en réponse aux risques systémiques identifiés par les plateformes sur leurs services et auxquels les *TTPs* contribuent (partie 2), il synthétise les mesures d'atténuation qu'elles ont prises pour réduire ceux-ci (partie 3). Cette partie est complétée d'une annexe qui identifie toutes les dispositions des conditions générales d'utilisation des plateformes, visant à prévenir les *TTPs*, et qui peuvent donc être mobilisées pour procéder à des signalements aux plateformes concernées afin de mieux protéger l'intégrité de nos espaces informationnels. Enfin, dans une dernière partie, l'Arcom formule quatre recommandations visant à amplifier la lutte contre les manipulations de l'information en réponse à l'accroissement des menaces, notamment du fait de la détérioration de l'environnement géopolitique.

Le cadre de corégulation entre les grandes plateformes et les régulateurs nationaux et européens a permis d'engager un dialogue structuré visant à répondre à l'exigence de lutte contre les manipulations de l'information, notamment à travers l'adoption des codes successifs contre la désinformation. Quant au RSN, il achève sa montée en puissance avec la création d'une capacité institutionnelle renforcée dans l'ensemble du réseau de régulation, l'activation graduelle des mécanismes de transparence imposés aux plateformes, la mise en œuvre d'un vaste programme de supervision et des premières actions de mise en conformité. Il s'accompagne de la mobilisation croissante de toutes les parties prenantes, notamment de la société civile et de l'ensemble des autorités publiques : gouvernements, parlements et autorités judiciaires.

Ces progrès permettent une montée en puissance de l'action publique, en cohérence avec la *Stratégie nationale de lutte contre les manipulations de l'information d'origine étrangère (2026-2030)* élaborée par le Secrétariat général de la Défense et de la Sécurité nationale (SGDSN) ; l'accroissement de la menace l'exige et invite aujourd'hui à adopter une approche plus prescriptive quant aux mesures attendues des grandes

plateformes pour lutter contre les techniques de manipulation de l'information utilisées par les acteurs malveillants.

Pour répondre à ces enjeux, l'Arcom recommande **le renforcement de l'effectivité de la mise en œuvre du RSN (recommandation n° 1)**, notamment la mise en œuvre rapide des dispositions relatives à :

- la transparence renforcée de la publicité numérique, avec l'**adoption de lignes directrices sur les registres publicitaires (article 39)** ;
- **l'accès aux données par les chercheurs agréés (article 40)**, afin de corriger les asymétries d'information entre les plateformes et nos démocraties et de documenter plus en profondeur l'évolution des risques systémiques et la pertinence des mesures d'atténuation de ces risques prises par les plateformes.

En outre, **l'adoption de lignes directrices en matière de lutte contre la manipulation de l'information** (en application de l'article 35 du règlement) permettrait de consolider le cadre issu du code de conduite et de préciser les attentes des autorités de régulation européennes et nationales à cet égard. L'Arcom souligne dans le même temps l'importance de **développer les mécanismes de coopération avec la société civile** mis en place par le code, qui constituent aujourd'hui des espaces d'échanges essentiels entre la Commission européenne, les plateformes signataires, les autorités de régulation nationales, les acteurs industriels et la société civile.

Toutefois, cette approche est conditionnée à l'existence de financements pérennes pour soutenir les travaux des vérificateurs de faits et de la recherche, qui participent à la préservation de l'intégrité de l'espace informationnel en ligne. **L'Arcom appelle donc les pouvoirs publics nationaux et la Commission européenne à créer les conditions permettant à ces acteurs de poursuivre leurs travaux, notamment à travers un soutien financier stable** (**recommandation n° 2**).

Il apparaît nécessaire de renforcer la coordination et le partage d'expertise et de bonnes pratiques entre les différents acteurs impliqués dans la lutte contre la manipulation de l'information en France, y compris hors période électorale et hors cas d'ingérence étrangère, pour inscrire ce travail dans la durée. À ce titre, **l'Arcom recommande la création d'un Observatoire sur la lutte contre les manipulations de l'information**, sur le modèle de l'Observatoire de la haine en ligne et des réunions de préparation des élections que l'Arcom organise avant chaque scrutin depuis 2024. Cet Observatoire aurait pour objectif d'inscrire l'effort collectif dans la durée et de mobiliser l'ensemble des acteurs sur des problématiques non-électorales (désinformation climatique par exemple) (**recommandation n° 3**).

Néanmoins, l'ensemble de ces dispositifs ne peuvent être pleinement efficaces sans une résilience informationnelle de l'ensemble des citoyens. Dans ce cadre, l'Arcom appelle de ses vœux **le renforcement et l'élargissement des actions d'éducation aux médias, à l'information et à la citoyenneté numérique à de nouveaux réseaux et publics, en dehors du cadre scolaire** (**recommandation n° 4**).

Introduction

Le législateur français a confié à l'Arcom la mission[1] de dresser un bilan des moyens et mesures mis en œuvre par les très grandes plateformes et les très grands moteurs de recherche en ligne[2] (les « ***grandes plateformes*** » ci-après, par facilité de langage) afin de lutter contre la manipulation de l'information sur leurs services.

En tant que coordinateur pour les services numériques (CSN) pour l'application du règlement européen sur les services numériques (RSN, *Digital services Act* ou DSA en anglais)[3] en France, l'Arcom est associée, aux côtés des CSN des autres États membres de l'Union européenne, au suivi des risques systémiques et des obligations renforcées pesant sur les grandes plateformes, dont la supervision est assurée par la Commission européenne[4].

Ce bilan présente l'état de la lutte contre la manipulation de l'information au sens du RSN, en étudiant les moyens mobilisés par les grandes plateformes pour lutter contre les tactiques, techniques et procédures (*TTPs*) utilisées par les acteurs malveillants pour perturber l'intégrité de leurs services, dans une approche qui ne porte pas sur les contenus et leur qualification. En effet, le RSN repose sur une logique de gestion des risques systémiques (articles 34 et 35 du règlement) et non sur une appréciation de la véracité des contenus pris isolément. Par conséquence, la lutte contre la « *désinformation* » – entendue comme l'identification et la qualification de contenus faux ou trompeurs – doit se distinguer de la lutte contre la « *manipulation de l'information* », qui vise prioritairement les comportements coordonnés ou les usages inauthentiques de fonctionnalités légitimes. Les *TTPs* constituent ainsi les vecteurs par lesquels les risques systémiques liés à la manipulation de l'information se matérialisent et sont amplifiés. Ces risques sont d'autant plus marqués en période électorale, les *TTPs* pouvant participer à « *tout effet négatif réel ou prévisible sur le discours civique, les processus électoraux et la sécurité publique* » (article 34 du RSN), et altérer la sincérité du scrutin et la confiance des citoyens envers les processus démocratiques et les institutions.

Dans ce contexte, les enjeux liés à la manipulation de l'information mobilisent aujourd'hui pleinement l'attention des pouvoirs publics, comme en témoigne la stratégie nationale de lutte contre les manipulations de l'information 2026-2030 du Secrétariat général de la défense et de la sécurité nationale, publiée le 11 février 2026[5] et portant sur le volet des ingérences numériques étrangères. À l'instar du présent bilan, cette stratégie privilégie une approche centrée sur les comportements et modes opératoires malveillants[6]. Elle met particulièrement en avant la nécessité de renforcer la régulation des plateformes en ligne, qu'elle identifie comme l'un des quatre piliers structurant l'action publique.

L'intégrité des services apparait ainsi comme l'une des composantes essentielles dans la lutte contre la manipulation de l'information. Les signataires du code de conduite contre la désinformation[7], dont l'intégration au RSN a pris effet le 1er juillet 2025, ont reconnu sa valeur en lui consacrant un chapitre dédié, visant à accroître l'engagement des grandes plateformes signataires face aux techniques de manipulation de

[1] Article 58 de la loi n° 86-1067 du 30 septembre 1986 relative à la liberté de communication modifié par l'article 54 de la loi n°2024-449 du 21 mai 2024 visant à sécuriser et réguler l'espace numérique (loi SREN).
[2] *Very large online platforms and search engines (VLOPSE)* en anglais.
[3] Règlement 2022/2065 du 19 octobre 2022 relatif à un marché unique des services numériques
[4] Annexe 2 : Actions de l'Arcom en 2025 en matière de lutte contre la manipulation de l'information en ligne.
[5] Stratégie nationale de lutte contre les manipulations de l'information, Secrétariat général de la défense et de la sécurité nationale, 11 février 2026.
[6] V. p.10, 3e paragraphe.
[7] https://disinfocode.eu/signatories

l'information communément identifiées. Sept grandes plateformes ont signé ce code de conduite - Facebook, Instagram, Google Search, YouTube, Microsoft Bing, Microsoft LinkedIn et TikTok – et ont souscrit des engagements *individuels spécifiques* de nature et de portée variable.

De ce fait, l'intégrité des services, qui fait appel à plusieurs notions complémentaires telles que la fiabilité et la robustesse des plateformes, l'authenticité, la résilience, la confidentialité ou encore la sécurité des grandes plateformes, concourt plus largement à maintenir un environnement numérique sûr et digne de confiance, et à assurer la qualité de l'information et la richesse et l'authenticité du débat public en ligne.

Néanmoins, l'action des plateformes visant à préserver l'intégrité de leurs services doit être conciliée avec le respect de la liberté d'expression, qui est un droit fondamental et constitue un des piliers de l'exercice démocratique. Le risque que la conception ou le fonctionnement de leurs services induisent des effets négatifs sur la liberté d'expression et d'information est d'ailleurs explicitement pré-identifié par le RSN[8]. En effet, une action trop stricte de leur part (conditions générales d'utilisation trop restrictives, sur-modération des contenus ou restriction à la création de compte non proportionnée par exemple) pourrait restreindre indûment la liberté d'expression, tandis qu'une approche trop permissive risquerait de laisser libre cours à des comportements inauthentiques ou malveillants. Le RSN intègre cet enjeu, en exigeant que les grandes plateformes adoptent des mesures d'atténuation des risques « *raisonnables, proportionnées et efficaces, adaptées aux risques systémiques spécifiques recensés conformément à l'article 34, en tenant compte en particulier de l'incidence de ces mesures sur les droits fondamentaux* » (article 35 du RSN).

Cette tension s'observe également lorsqu'il s'agit de concilier la sécurité des plateformes, qui implique de ne pas révéler de manière détaillée leurs modalités de fonctionnement, et les obligations de transparence imposées aux grandes plateformes[9] en application du RSN. De ce fait, la lutte contre les *TTPs* doit être conçue de manière à concilier les enjeux de sécurité et de fiabilité des services, et les libertés publiques.

Ce bilan propose donc de mettre en perspective la typologie des *TTPs* issue du code de conduite contre la désinformation avec les risques systémiques identifiés par les grandes plateformes sur leurs services. Il expose également les mesures d'atténuation mises en œuvre par ces derniers pour limiter l'impact des *TTPs* et fournit un ensemble de propositions, destinées tant aux pouvoirs publics qu'aux grandes plateformes et aux utilisateurs, visant à renforcer l'intégrité des services et prévenir la manipulation de l'information par une amplification artificielle de la diffusion de certains contenus.

Cet exercice impose de porter une attention particulière à l'écart potentiel entre les déclarations des grandes plateformes, les dispositifs qu'elles affirment déployer et l'efficacité concrète de ces mesures face à des techniques de manipulation de l'information encore pleinement observables et en constante évolution sur leurs services. Sans se limiter à une approche purement déclarative, l'Arcom a choisi de s'appuyer, pour son analyse, sur l'ensemble des ressources utiles dont elle dispose, notamment les travaux du service de vigilance et de protection contre les ingérences numériques étrangères (VIGINUM), des organisations de la société civile ainsi que de

[8] Article 34, 1. b)

[9] Par exemple, concernant l'intelligence artificielle, Meta explique que « *la transparence et la sécurité sont parfois difficiles à conjuguer. Révéler en détail le fonctionnement de certains systèmes d'IA pourrait compromettre les systèmes de sécurité ou ouvrir la voie à des attaques adverses, et donc porter potentiellement préjudice aux personnes qui utilisent nos produits […]. Mais d'un autre côté, il est important pour nous d'informer nos utilisations sur le fonctionnement de nos systèmes d'IA. Nous cherchons donc à atteindre le bon équilibre en matière de transparence* ». (https://about.fb.com/fr/news/2022/02/les-cartes-de-systeme-une-nouvelle-ressource-pour-comprendre-le-fonctionnement-des-systemes-dia/)

chercheurs. Ce croisement de sources et d'expertises permet de mieux apprécier la pertinence des moyens mis en œuvre par les grandes plateformes.

Avertissement

Les grandes plateformes étudiées pour ce bilan sont : Facebook et Instagram (Meta), Google Search, YouTube, Microsoft Bing, LinkedIn, X, TikTok, Snapchat et Wikipédia.

L'analyse se fonde sur les éléments et données figurant :

- dans les rapports d'évaluation et d'atténuation des risques systémiques publiés en 2024 et en 2025 en application du RSN (articles 34, 35 et 42) ;
- dans les rapports d'audit (article 37 du RSN) ;
- dans les rapports publiés en mars et en septembre 2025 en application du code de conduite contre la désinformation ;
- dans les clauses pertinentes des conditions générales d'utilisation des services des grandes plateformes.

Les prises de positions de ces derniers, issues de leurs rapports publics et présentées dans ce bilan, n'engagent pas l'Arcom. Leurs mentions ne sauraient être interprétées comme une approbation du régulateur, dont les analyses dans le présent bilan sont sans préjudice de l'évaluation des obligations d'évaluation des risques systémiques relevant de la Commission européenne.

Les observations figurant dans le présent bilan s'apprécient à la date de sa publication et ne préjugent pas des évolutions susceptibles d'être apportées par les grandes plateformes à la conception et au fonctionnement de leurs services.

PARTIE 1. TYPOLOGIE DES TECHNIQUES DE MANIPULATION DE L'INFORMATION SUR LES PLATEFORMES EN LIGNE

Dans un environnement numérique où circulent quotidiennement des volumes massifs de contenus, le risque de manipulation de l'information constitue un enjeu central pour les grandes plateformes et leurs utilisateurs. Les acteurs malveillants mobilisent des techniques de plus en plus variées et sophistiquées, exploitant aussi bien les vulnérabilités des services que les erreurs humaines.

Dans ce contexte, les signataires du code de conduite contre la désinformation (ci-après, le code)[10] ont reconnu l'importance d'intensifier leurs efforts pour garantir l'intégrité des services, en mettant en œuvre des mesures de protection contre les comportements malveillants ou inauthentiques et la diffusion de la désinformation. Soulignant l'importance de l'approche collaborative (entre plateformes mais aussi avec la société civile), ils ont élaboré une liste commune de *TTPs*, qui doit être périodiquement mise à jour à la lumière des évolutions technologiques et des preuves illustrant de nouvelles menaces.

Même si l'Arcom considère que cette liste est aujourd'hui en partie devenue incomplète[11], elle permet toutefois de fournir une grille d'analyse utile qui inscrit les techniques identifiées dans un continuum allant de la création à la diffusion de la désinformation, en passant par le ciblage. Elle donne également un premier aperçu de l'impact de l'innovation technologique, notamment de l'intelligence artificielle (IA) générative, parfois utilisée dans le cadre de *TTPs* pour créer et diffuser de façon accélérée et massive des contenus véhiculant de la désinformation et pour contourner les mécanismes interdisant de créer et d'automatiser de faux comptes. Ce phénomène a par ailleurs donné lieu à la création du concept de « *Slopaganda* »[12], ce mot-valise mêlant l' « *AI Slop* » (c'est-à-dire des contenus jetables, de faible qualité, générés par l'IA) et la notion de propagande, pour désigner ces contenus diffusés à grande échelle visant à influencer les opinions et les croyances, à des fins politiques ou idéologiques[13].

Cette liste commune de *TTPs* est associée à une série d'engagements pris par les signataires en faveur de mesures de transparence concernant la protection de l'intégrité de leurs services. Dans ce cadre, ils doivent notamment rendre compte au public des mesures mises en œuvre pour lutter contre les techniques de manipulation de l'information sur leurs services, afin que ces dispositions puissent faire l'objet d'une évaluation de leur pertinence et de leur impact éventuel sur la liberté d'expression.

Toutefois, la défense de l'intégrité des services demeure délicate et ne pourra être menée à bien sans des efforts importants et continus des signataires, dans la mesure où les acteurs malicieux détournent des fonctionnalités légitimes des plateformes. Les CGU des grandes plateformes proscrivent donc certains usages malveillants de leurs services et ces restrictions doivent être revues périodiquement pour tenir compte des nouvelles formes d'abus par des acteurs malveillants. Ainsi, par exemple, plusieurs

[10] À sa création en 2018, ce code était initialement un code de bonnes pratiques de lutte contre la désinformation. Il a été renforcé en 2022 et converti en code de conduite en juillet 2025 au sens de l'article 45 du RSN, ce qui rend les engagements opposables aux grandes plateformes signataires.
[11] V. partie 4 – Recommandations.
[12] Ce concept a été théorisé par les chercheurs Michał Klincewicz, Mark Alfano et Amir Fard : https://repository.tilburguniversity.edu/server/api/core/bitstreams/07b4d52f-4cef-451a-ad98-536082cd224b/content
[13] La « *Slopaganda* » exploite les dynamiques des plateformes en ligne et des moteurs de recherche pour saturer l'environnement informationnel, créer un « *bruit* » médiatique et submerger les utilisateurs – ce qui peut affaiblir la vigilance, favoriser certaines croyances ou plus généralement participer à une manipulation de l'information.

plateformes interdisent l'utilisation de bots ou de scripts pour écrire de faux avis ou commentaires, ou pour augmenter le nombre de « *J'aime* » ou les partages.

I. Création de l'inauthenticité

A. Les comportements inauthentiques coordonnés

Les comportements inauthentiques coordonnés, traduite de l'anglais « *Coordinated Inauthentic Behavior* » (CIB)[14], font l'objet de nombreux développements dans les rapports des grandes plateformes signataires du code de conduite contre la désinformation[15]. La notion regroupe une série d'actions menées de façon organisée et dissimulée par des acteurs malveillants aux motivations variées (opération d'influence, volonté de déstabilisation, motivation idéologique, gain financier, etc.).

Par exemple, la création de faux comptes (« *sock-puppets*[16] ») ou la mise en réseau de comptes inauthentiques et automatisés (« *botnets*[17] »), articulés de manière inauthentique, permet de façonner artificiellement l'apparence d'un consensus, d'influencer la visibilité des contenus et de polariser le débat public. Ces comptes fictifs et automatisés peuvent poster, partager, *liker*, commenter des contenus, acheter des faux engagements et suivre massivement d'autres comptes[18].

Concrètement, ces *TTPs* fonctionnent selon plusieurs leviers complémentaires :
- les faux comptes s'intègrent dans des communautés en ligne et adoptent des comportements proches de ceux d'utilisateurs réels (biographie, photo de profil, interactions, etc.) ;
- ils multiplient les créations de contenus de faible qualité, réagissent et partagent massivement ces derniers, afin de manipuler les algorithmes de recommandation des plateformes et de leur donner une résonance disproportionnée ;
- leur engagement est coordonné et ciblé : ils peuvent viser des utilisateurs influents ou des communautés vulnérables pour accroître l'impact de leurs contenus, attirer l'attention des utilisateurs voire façonner progressivement leur opinion.

Quelques signaux faibles permettant de détecter un faux compte ou un réseau inauthentique

Signaux liés au profil du compte :
- Une usurpation d'identité (notamment d'une personnalité publique) ;
- Un nom d'utilisateur d'apparence inauthentique : nom mélangeant des mots étranges, contenant des chiffres aléatoires ou suivant un modèle peu naturel (ex. « CTUGDF35505 ») ;
- Des informations de profil (ex. localisation, date de création du compte) incohérentes vis-à-vis du contenu publié.

[14] Notion employée par Meta dans ses rapports de transparence.
[15] Les grandes plateformes signataires du code de conduite contre la désinformation sont les suivants : Google Search, YouTube, Facebook, Instagram, TikTok, Microsoft Bing, LinkedIn.
[16] Faux comptes utilisés à des fins trompeuses.
[17] Automatisés, partiellement automatisés ou non automatisés.
[18] Cela prend en compte les comptes fantômes (« *ghost accounts* »), c'est-à-dire des comptes inactifs ou passifs, utilisés pour gonfler les statistiques.

Signaux comportementaux :

- Des publications très fréquentes, massives ou synchronisées avec d'autres comptes ;
- Des copier-coller de blocs de textes similaires sur une ou plusieurs plateformes, dans une même temporalité (technique dite du « *copy-pasta* ») ;
- Des partages et réactions systématiques ;
- Des contenus monothématiques, polarisants ou promotionnels.

À noter que certaines plateformes interdisent explicitement la création de faux comptes inauthentiques, qu'ils soient exploités sous pseudonymes ou qu'ils procèdent à l'usurpation de l'identité d'un tiers (Meta, X, Microsoft Bing, LinkedIn). De façon plus globale, toutes les plateformes étudiées prohibent l'usurpation d'identité sur leurs services.

En cas de doute concernant un compte dont l'activité serait susceptible d'aller à l'encontre des CGU d'une plateforme ou du cadre juridique en vigueur, il est possible de le signaler au service concerné, en tenant compte du fait qu'un pseudonyme ou certains signaux isolés ne signifient pas nécessairement que le compte est inauthentique, et donc que la plateforme appréciera de la pertinence du signalement et agira en conséquence.

B. La création de pages, de groupes de conversations, de forums

Créer des pages, des groupes de conversation ou des forums d'échange permet également de bâtir des communautés factices et donne l'illusion d'un lieu de discussion légitime, souvent fréquenté et crédible. Ce sont des espaces actifs qui servent en réalité d'infrastructures de propagation. Ces entités pouvant publier ou relayer des contenus trompeurs ou polarisants, avec l'apparence d'une production collective ou citoyenne. Leurs objectifs sont de simuler l'activité humaine, d'influencer les opinions, de diffuser de la désinformation ou de créer des engagements inauthentiques.

Exemples :

- *Une page prétendant être celle d'une célébrité pour se prévaloir de sa notoriété, attirer ainsi des abonnés et générer de l'engagement ;*
- *Une page publiant des résultats de faux sondages pour influencer l'opinion publique.*

C. La création de noms de domaines inauthentiques

Également perçus comme des nœuds de réseau coordonné, les noms de domaine inauthentiques sont créés pour tromper sur l'identité, l'origine ou la légitimité d'un site web. Autrement dit, l'objectif est d'usurper l'identité d'une organisation ou d'une personne légitime, en intégrant de légères variations.

Ses caractéristiques principales sont les suivantes :

- une orthographe ou une graphie trompeuse (« *typosquatting* ») : des caractères d'apparence proche (lettres latines, cyrilliques ou des chiffres) sont utilisés pour donner l'impression d'un nom valide lors d'une lecture rapide (ex. https://www.arcorn.fr/ au lieu de https://www.arcom.fr/). Dans cet exemple, la

forme des lettres est altérée pour tromper l'œil du lecteur : le premier lien comporte un « r » et un « n » juxtaposés, pouvant être perçus comme un « m » et donnant l'illusion que le mot affiché est « arcom ». Un utilisateur pourrait ainsi croire visiter le site officiel, alors qu'il s'agit d'un nom de domaine frauduleux ;
- une différence d'extension de nom de domaine (« *top-level domain* » ou « TLD ») : les acteurs malveillants peuvent choisir une extension moins visible ou moins attendue que l'originale. Par exemple, le site légitime pourrait être en « .fr » ou « .com », alors que le faux domaine utilisera « .net », « .info », pour semer la confusion (ex. https://www.arcom.net/ au lieu de https://www.arcom.fr/).

Ces différentes formes de tromperie misent sur le manque de vigilance des utilisateurs et visent à créer un sentiment de familiarité ou de légitimité, un nom proche d'un site connu ou une extension plausible pouvant leur donner confiance.

D. Le détournement de compte ou l'usurpation d'identité

Le détournement de compte (« *hijacking* ») ou l'usurpation d'identité sont des pratiques qui relèvent de l'ingénierie sociale – un ensemble de techniques de manipulation, « *consistant à obtenir un bien ou une information, en exploitant la confiance, l'ignorance ou la crédulité de tierces personnes* »[19].

Ces techniques permettent de donner une apparence de légitimité à des contenus ou des messages, ce qui les rend plus crédibles, même s'ils sont faux ou manipulés. Combinées à d'autres *TTPs* (cf. *supra*), ces pratiques deviennent un levier puissant pour influencer le débat et orienter les opinions, tout en rendant plus difficile la détection et la modération.

- Dans le cas d'un compte détourné, c'est-à-dire piraté (en anglais « *Account Takeover Fraud* » - ATO), l'acteur malveillant accède sans autorisation aux informations d'identification du compte d'un utilisateur et réussit à en prendre le contrôle[20]. Il agit donc sous l'identité de sa victime, accède à ses informations, manipule son réseau de contacts et diffuse des contenus ou des messages.
- Dans le cas de l'usurpation d'identité (ou création d'une identité falsifiée), l'acteur malveillant créé de toutes pièces, ou à partir de données volées, une fausse identité.

Ces deux procédés représentent des menaces concrètes pour l'intégrité des plateformes en ligne, en ce qu'ils engendrent une perte de contrôle sur l'identité d'une personne, qu'elle soit physique ou morale, et peuvent également servir de relais crédibles pour diffuser de fausses informations.

L'ANSSI fournit « *10 règles d'or en matière de sécurité numérique* » pour se prémunir contre le détournement de comptes en ligne. Ces recommandations visent à réduire ou compliquer les voies d'accès qu'un acteur malveillant pourrait exploiter (mot de passe faible, *phishing*, logiciel malveillant, etc.). Elles participent à la protection de la confidentialité, de l'intégrité, de la disponibilité et de l'authenticité des comptes des utilisateurs.

[19] Définition de l'ANSSI : https://cyber.gouv.fr/le-cyberdico
[20] Il peut le faire via diverses techniques : attaque par hameçonnage, récupération de mot de passe, bourrage d'identifiants (« *credential stuffing* »), i.e. utilisation massive de combinaison identifiants/mots de passe volés, etc.

En conclusion, les conséquences liées à la mise en œuvre de ces *TTPs* sont multiples et peuvent être d'une gravité importante, notamment dans un contexte électoral.

D'une part, elles favorisent la création massive de contenus de désinformation, ce qui peut affaiblir la confiance collective. En effet, lorsqu'un grand nombre d'utilisateurs constate que des engagements, des soutiens populaires ou encore des avis ne sont en réalité que des contenus artificiels, la crédibilité accordée aux contenus authentiques des médias, des personnalités ou des institutions peut s'en trouver réduite par ricochet.

D'autre part, cette dynamique favorise la polarisation et la fragmentation de l'espace public, ces réseaux inauthentiques créant des « *chambres d'échos* » (autrement appelées « *bulles informationnelles* »), qui captent l'attention de certains utilisateurs.

Enfin, ces *TTPs* peuvent permettre la manipulation de l'agenda politique : en donnant l'illusion d'un soutien massif ou d'un mécontentement large, elles peuvent orienter le débat et influencer certaines décisions politiques voire électorales.

II. Diffusion et amplification de la désinformation

Afin d'optimiser leurs effets, les *TTPs* visant à créer des moyens inauthentiques et des contenus véhiculant de la désinformation (cf. *supra*) s'accompagnent bien souvent de techniques de diffusion et d'amplification visant à partager ces contenus au-delà de cercles restreints, pour en faire un phénomène massif.

A. Le ciblage de publics vulnérables à travers la publicité en ligne

- ***La manipulation de l'opinion***

Le ciblage consiste à segmenter les utilisateurs en fonction de leurs caractéristiques (âge, situation sociale, habitudes en ligne, intérêts, etc.), pour leur envoyer des contenus qui peuvent attirer leur attention. Par l'usage de leurs services, les utilisateurs fournissent aux plateformes des quantités importantes d'informations personnelles (sur leurs préférences, leur localisation, leur vie privée, etc..). Via leur régie publicitaire, les grandes plateformes permettent à leurs annonceurs de mobiliser ces informations pour procéder à des ciblages de groupes d'individus en vue de les exposer à leur publicité ou autres contenus sponsorisés.

Le ciblage publicitaire peut être utilisé à des fins commerciales afin d'influencer les comportements de consommation des utilisateurs, par exemple en les incitant à acheter un produit ou à utiliser un service.

Cette logique de ciblage s'applique également dans le domaine politique : i.e. contrairement à la publicité commerciale, la publicité politique ne vise pas à promouvoir un service ou un produit, mais à mettre en avant une information ou une opinion, qui peut être utilisée pour promouvoir des candidats, des partis ou des idées, en jouant éventuellement sur les biais cognitifs des utilisateurs.

Dans le cadre de campagnes de manipulation de l'information, ce sont souvent les publics les plus vulnérables qui sont ciblés *via* la publicité ; leurs fragilités et leurs émotions (peur, colère, incertitude, etc.) étant exploitées pour maximiser l'efficacité de l'opération d'influence. Cette technique peut s'accompagner de moyens de dissimulation de l'identité réelle de l'auteur de l'opération, via l'utilisation d'adresses IP usurpées (*IP*

spoofing) ou de techniques d'obscurcissement[21], qui peuvent servir à masquer l'origine réelle du ciblage et à rendre l'action difficile à tracer ou à attribuer.

Cadre juridique applicable à la publicité politique

La publicité politique en ligne fait l'objet d'encadrements juridiques distincts au niveau national et européen, qui se cumulent mais diffèrent dans leur champ d'application et dans les restrictions imposées. Le droit national appréhende la publicité politique à travers une notion étroite et très encadrée, assortie d'interdictions d'usages substantielles, tandis que le droit européen en retient une définition très large mais impose des règles de transparence et de traçabilité ainsi que des restrictions de ciblage, sans interdiction d'usage.

Au niveau national, l'article 14 de la loi n° 86-1067 du 30 septembre 1986 relative à la liberté de communication pose une interdiction générale et permanente des « *émissions publicitaires à caractère politique* » à la télévision et à la radio. Par ailleurs, l'article L. 52-1 du code électoral prévoit que « *pendant les six mois précédent le premier jour du mois d'une élection et jusqu'à la date du tour de scrutin où celle-ci est acquise, l'utilisation à des fins de propagande électorale de tout procédé de publicité commerciale par la voie de presse ou par tout moyen de communication audiovisuelle est interdite* ». La propagande électorale correspond à l'ensemble de la communication à laquelle les candidats ont recours pour faire campagne, notamment les affiches, professions de foi, tracts, etc.

Au niveau européen, le règlement relatif à la transparence et au ciblage de la publicité à caractère politique[22], entré en vigueur le 10 octobre 2025, retient une notion très large de la publicité politique : toute publicité conçue dans le but d'influencer les résultats d'une élection ou d'un référendum, un comportement de vote ou un processus législatif ou réglementaire au niveau de l'UE. L'approche retenue par le règlement vise plutôt à imposer des obligations renforcées en matière de transparence ou de restriction de ciblage : par exemple, chaque annonce publicitaire à caractère politique doit faire l'objet d'un marquage clair (article 11). Les prestataires de services de publicité à caractère politique doivent tenir un registre publicitaire permettant d'assurer une traçabilité et un suivi des publicités mises en ligne (article 9) « *pendant une période de sept ans à compter de la date de l'ultime élaboration, placement, promotion, publication ou diffusion, selon le cas* ». Les possibilités et les modalités de recourir au ciblage publicitaire sont encadrées (articles 18 et 19).

À noter qu'un projet de loi portant diverses dispositions d'adaptation au droit de l'Union européenne en matière économique, financière, environnementale, énergétique, d'information, de transport, de santé, d'agriculture et de pêche (DADDUE 2026) est actuellement en cours de discussion au Parlement et tend à adapter le droit national au règlement susmentionné. L'Arcom serait désignée autorité compétente pour l'application en France de la plupart des dispositions du règlement, aux côtés de la Commission nationale de l'informatique et des libertés (CNIL).

[21] Ces techniques consistent à transformer du code, des données, du trafic réseau ou des communications de manière délibérée pour les rendre difficiles à retracer et à analyser. Autrement dit, l'obscurcissement ne cherche pas à bloquer l'accès (comme le fait le chiffrement de données), mais à cacher ou dissimuler la structure, l'origine ou la logique d'un contenu ou d'une action.
[22] Règlement (UE) 2024/900 du 13 mars 2024 relatif à la transparence et au ciblage de la publicité à caractère politique : https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=OJ:L_202400900

Focus sur le microciblage

Le microciblage désigne une stratégie de communication consistant à exploiter des données personnelles ou comportementales très fines d'individus ou de petits groupes, pour segmenter les publics en groupes restreints et leur envoyer des messages personnalisés. Il s'agit donc de construire des bulles sur-mesure, chaque segment (ou individu) recevant une version du contenu qui maximise les chances qu'il le « *reçoive bien* », c'est-à-dire, par exemple, en résonance avec ses peurs, ses préoccupations, ses valeurs, son style de vie, etc.

*Exemple : le rapport d'enquête[23] de l'*Information Commissioner's Office (ICO)[24], *portant sur l'utilisation des données dans les campagnes politiques, revient sur les pratiques de la société Cambridge Analytica lors de la campagne présidentielle américaine de 2016, notamment l'exploitation de données issues de réseaux sociaux pour construire des profils d'électeurs, afin d'envoyer des publicités politiques personnalisées.*

- ***La désinformation à finalité lucrative***

Les campagnes de manipulation de l'information constituent une source directe de revenus pour certains acteurs malveillants et la recherche de profits est parfois la motivation principale, si ce n'est exclusive, de certaines campagnes de manipulation de l'information.

Il est ainsi possible d'observer que les modèles économiques des plateformes peuvent permettre à la désinformation d'être rémunératrice à plusieurs titres :

- **à travers les escroqueries** : dans ce cadre, la diffusion de fausses informations sous la forme d'un message publicitaire vise directement à tromper les utilisateurs pour obtenir de façon illicite leur argent, *via* des produits, services ou transactions frauduleuses. Ces contenus, souvent présentés comme fiables et légitimes, exploitent la crédulité ou la confiance des consommateurs, parfois de manière très sophistiquée. Ils peuvent se caractériser également par une dimension sensationnaliste destinée à capter l'attention et à inciter au clic, en jouant sur la curiosité ou l'émotion des utilisateurs. Dans ce système, la rentabilité est rapide et directe, chaque victime constituant un gain potentiel pour l'acteur malveillant ;
- **à travers la monétisation de l'attention** : dans ce contexte, c'est l'attention des utilisateurs qui devient la source de revenus sous la forme d'un intéressement de l'auteur des contenus à la recette publicitaire versée par l'annonceur. Les contenus, souvent sensationnalistes et trompeurs, sont conçus pour maximiser l'engagement des utilisateurs et donc leur exposition aux publicités qui accompagnent ces contenus. La plateforme, qui commercialise ces publicités, tire elle aussi profit de la diffusion massive de ces contenus trompeurs. Ces contenus peuvent également comporter des liens renvoyant vers des sites tiers intégrant de la publicité non commercialisée par la plateforme, augmentant ainsi le gain de ces acteurs malveillants.

Ces cas d'usage démontrent que le modèle économique des plateformes en ligne peut favoriser la dynamique économique de la désinformation, tout particulièrement à travers

[23] https://ico.org.uk/media2/migrated/2260271/investigation-into-the-use-of-data-analytics-in-political-campaigns-final-20181105.pdf
[24] Autorité britannique chargée de la protection des données et de la vie privée.

leurs algorithmes de recommandation et leur système de monétisation. La nécessité de maitriser ce risque est identifié comme l'un des objectifs stratégiques de la Stratégie nationale de lutte contre les manipulations de l'information 2026-2030. L'Arcom considère que ces exemples devraient conduire les plateformes à redoubler de vigilance et à développer des moyens adaptés pour détecter et agir contre les acteurs malveillants qui développent des contenus relevant de la désinformation à des fins purement lucrative.

L'intensité de ces pratiques est susceptible d'être décuplée par l'usage des intelligences artificielles génératives, qui facilitent la création, la traduction et la diffusion massive de contenus trompeurs.

B. L'utilisation de contenus truqués et/ou trompeurs (*deepfakes*)

Un hypertrucage ou « *deepfake* » est un « *trucage audio ou vidéo à partir d'éléments existants, utilisant l'intelligence artificielle pour changer le visage d'une personne dans une vidéo ou reproduire sa voix* »[25]. Autrement dit, ce type de contenu fait dire ou fait faire à une personne (ex. personnalité politique, expert, etc.) des propos qu'elle n'a jamais tenus ou des actes qu'elle n'a jamais commis. Cette technique peut servir à diffuser des accusations fausses, des récits trompeurs, ou nuire à une personne.

L'utilisation de *deepfakes* en période électorale est particulièrement dangereuse, notamment pour les raisons suivantes :

- elle peut influencer l'opinion ou le comportement des électeurs, en les poussant par exemple à voter pour ou contre un candidat, ou en les incitant à l'abstention, remettant ainsi en cause la sincérité du scrutin ;
- la rapidité de leur diffusion et leur caractère massif peuvent nuire à la capacité de vérification des faits, qui peut intervenir trop tardivement pour démystifier certains contenus truqués ;
- elle participe à fragiliser la confiance collective, l'utilisation de l'intelligence artificielle pouvant amener les citoyens à « *douter de tout* », ce qui affaiblit le débat démocratique et rend difficile la formation d'une conviction libre et éclairé.

Exemple : lors des élections fédérales allemandes de février 2025, l'opération « Storm-1516 », liée aux tentatives d'ingérence russe, a utilisé l'intelligence artificielle pour créer plus d'une centaine de sites web, ayant relayé de nombreux deepfakes *visant des personnalités politiques comme Annalena Baerbock, Robert Habeck et Marcus Faber*[26]*.*

Comment détecter les contenus générés par l'IA ?

Il est possible de s'appuyer sur les étiquetages ou mentions ajoutées par certaines plateformes, qui indiquent explicitement qu'un texte, une image ou une vidéo est issu (totalement ou partiellement) d'une IA. Toutefois, ces étiquettes ne sont pas toujours présentes ou fiables ; la décision effective d'étiqueter un contenu relevant souvent de la responsabilité de l'usager, qui publie le contenu, même si la plateforme l'impose via ses CGU.

[25] VIGINUM, Guide de sensibilisation à l'attention des équipes de campagne (2025).
[26] https://belux.edmo.eu/fr/lections-fdrales-allemandes-la-dsinformation-et-les-tentatives-dingrences-pseront-elles-sur-les-rsultats/

En complément, AI Forensics a publié un guide comportant une série d'étapes que l'utilisateur peut suivre pour évaluer la probabilité que le contenu qu'il consulte ait été généré à l'aide d'outils d'IA.

C. Les opérations d'intrusion informatique suivie de fuites de données (« *hack and leak* »)

L'intrusion informatique suivie d'une fuite de donnée volontaire ou « *hack and leak* » est une technique visant à pénétrer (« *hack* ») au sein d'un système informatique, afin d'y dérober des informations sensibles ou privées, puis de les rendre publiques (« *leak* »). Cette technique, qui s'inscrit dans une logique d'exploitation informationnelle, est souvent utilisée pour influencer l'opinion publique, nuire à une personne morale ou physique, exercer une pression politique ou économique, ou encore déstabiliser une institution.

Les enjeux liés aux fuites de données orchestrées sont multiples et dépassent largement la seule dimension technologique. D'un point de vue sécuritaire, cette technique expose les personnes physiques et morales visées à la divulgation non autorisée de données personnelles ou d'informations stratégiques sensibles, entraînant ainsi leur compromission. Sur le plan réputationnel, elle peut provoquer une perte durable de confiance auprès des citoyens, en raison de l'exposition médiatique des informations.

Dans un contexte politique, les fuites de données volontaires sont un outil de manipulation particulièrement puissant : elles peuvent influencer des processus électoraux, fragiliser les institutions ou servir de leviers dans les conflits internationaux.

Exemple : les « Macron Leaks », survenus deux jours avant le second tour de l'élection présidentielle française de 2017, ont été caractérisés par la diffusion massive d'un ensemble de documents présentés comme provenant de l'équipe de campagne d'Emmanuel Macron. Ces informations ont été obtenues grâce à une opération de piratage[27] informatique de certains comptes personnels et professionnels.

D. Exemples de techniques de coordination inauthentique destinées à la création et à l'amplification des contenus

Les techniques de coordination inauthentiques destinées à créer et amplifier des contenus regroupent un ensemble de *TTPs* utilisées pour donner artificiellement l'impression qu'une information, une opinion ou un récit bénéficie d'un soutien massif. Ces *TTPs* reposent souvent sur des réseaux de comptes coordonnés (cf. *supra*), des automatisations ou des intermédiaires rémunérés.

Exemple : lors de l'élection présidentielle roumaine de novembre 2024, des influenceurs ont été recrutés pour publier des vidéos ou des messages de soutien au candidat Călin Georgescu, parfois via des scripts fournis par une agence de communication. Cette méthode s'inscrit plus largement dans une vaste opération d'ingérence étrangère, qui a conduit à l'annulation du premier tour de l'élection par la Cour constitutionnelle roumaine.

[27] L'opération de piratage informatique peut s'entendre comme une série d'actions visant à accéder, exploiter ou perturber un système ou un réseau informatique, notamment dans l'objectif d'obtenir un accès non autorisé à des informations sensibles ou privées, comme des données personnelles.

Parmi les pratiques courantes, figure le **bourrage par mots-clés**, qui consiste à publier de très nombreux contenus comportant les mêmes *hashtags* ou termes, afin de faire remonter un sujet dans les tendances et de donner l'illusion qu'il s'agit d'un phénomène spontanément populaire. Si cette technique peut être réalisée de manière automatisée ou semi-automatisée, elle peut également s'appuyer sur des **influenceurs**, conscients ou non de leur participation à l'opération d'influence.

Les vides de données ou « ***data voids »*** font également partie des techniques de manipulation des services qui visent à exploiter leurs algorithmes de recommandation pour amplifier artificiellement certains contenus. Elles permettent de fausser les indicateurs que les algorithmes interprètent comme des preuves de popularité ou de pertinence, pour mettre en avant certaines informations. Cette technique permet d'exploiter une zone de vide informationnel, c'est-à-dire de cibler des mots-clés ou des sujets présentant très peu de contenus indexés par les grandes plateformes, afin que les contenus trompeurs apparaissent en tête des résultats lorsque les utilisateurs cherchent ces termes. Ces informations fausses ou trompeuses deviennent ainsi la source par défaut des utilisateurs, puisqu'il n'y a pas d'autres informations fiables et pertinentes. Les vides de données ou « *data voids* » sont particulièrement efficaces lors de crises, d'élections ou d'événements soudains, dans le cadre desquels de nouveaux termes émergent avant que les médias ou les experts ne puissent produire des analyses fiables. Dans ces circonstances, les acteurs qui occupent rapidement ces espaces de vide informationnel peuvent orienter la perception des utilisateurs, façonner la compréhension d'un sujet, et imposer un récit biaisé qui sera repris et amplifié.

E. La sponsorisation dissimulée de contenus diffusés par des influenceurs

Lorsqu'un utilisateur influent publie un contenu sans indiquer qu'il est sponsorisé (c'est-à-dire qu'il reçoit une rémunération en retour de sa communication au public), les utilisateurs, notamment ses abonnés, peuvent le percevoir comme une recommandation personnelle ou une opinion spontanée, ce qui peut augmenter son impact.

À noter que la loi du 9 juin 2023 visant à encadrer l'influence commerciale et à lutter contre les dérives des influenceurs sur les réseaux sociaux[28], dans sa version en vigueur modifiée par l'ordonnance n° 2024-978 du 6 novembre 2024, vise à réduire ces pratiques dissimulées en favorisant une meilleure transparence permettant aux utilisateurs d'identifier les contenus sponsorisés et diffusés dans le cadre d'activités d'influence commerciale.

Dans le même sens, et en application de l'article 26 du RSN, les plateformes en ligne doivent fournir à leurs utilisateurs (y compris les influenceurs commerciaux) une fonctionnalité leur permettant de déclarer la diffusion de communications commerciales au moyen de « *marquages bien visibles* ». Les destinataires du service doivent ainsi pouvoir « *de manière claire, précise, non ambiguë et en temps réel* », avoir accès à des informations détaillées sur le contenu (sponsorisation, personne pour le compte de laquelle la publicité est présentée, personne l'ayant payée, ainsi que toute information utile concernant les principaux paramètres utilisés pour cibler les utilisateurs).

Cette technique peut être exploitée pour orienter des comportements électoraux, favoriser ou marginaliser certaines opinions, tout en donnant l'impression aux

[28] https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000047663185

utilisateurs que le soutien est organique et non publicitaire (cf. élections présidentielles roumaines de 2024, mentionnées *supra*).

F. Le signalement massif et coordonné de comptes ou de contenus

Cette technique consiste à abuser des outils légitimes de signalement mis à disposition des utilisateurs pour saturer les systèmes de modération, par un afflux soudain et volumineux de signalements. Si les grandes plateformes peuvent restreindre les capacités de signalement des acteurs malveillants, cette technique peut également s'apparenter à une attaque par déni de service (attaque « *DDos – Distributed Denial of service* »), qui porte atteinte à l'intégrité des grandes plateformes et conduire à des conséquences diverses :

- une suspension temporaire de l'accès à leurs outils de signalement pour tous les utilisateurs, liée à un dysfonctionnement résultant de l'effet de saturation ;
- la suppression de contenus ou de comptes pourtant conformes aux conditions d'utilisation des grandes plateformes et au cadre juridique en vigueur.

Ce détournement de fonctionnalités créé un effet d'étranglement : il ralentit la gestion des signalements légitimes et fragilise la fiabilité des systèmes de modération.

***

En conclusion, cette typologie des *TTPs* met en évidence la diversité et la sophistication des techniques de manipulation de l'information sur les plateformes en ligne. Reposant sur une combinaison de coordination humaine et d'automatisation, elles exploitent à la fois la crédulité des utilisateurs et les fonctionnalités légitimes offertes par les grandes plateformes (création de compte, engagements – *likes*, partages, commentaires, abonnements – outils d'intelligence artificielle, dispositif de signalement, etc.). Les grandes plateformes elles-mêmes reconnaissent que ces *TTPs* représentent des risques importants pour la sécurité, la fiabilité et la qualité de l'information sur leurs services. Pour autant, l'Arcom regrette que ceux d'entre eux qui sont signataires du code de conduite contre la désinformation ne fournissent pas de données chiffrées comparables concernant la détection de telles techniques, qui ne permettent pas d'avoir un état des lieux clair de l'état de la menace.

III. Présentation du concept de mode opératoire informationnel par VIGINUM

L'Arcom coopère étroitement avec VIGINUM depuis la création de ce service en 2021. VIGINUM est chargé de lui fournir toute information utile dans l'accomplissement des missions qui lui sont confiées par la loi du 30 septembre 1986 relative à la liberté de communication[29] et la loi du 21 juin 2004 pour la confiance dans l'économie numérique[30].

Afin de faciliter le partage d'information avec le service et sa contribution à la détection des risques systémiques en matière d'ingérence numérique étrangère, une convention cadre de partenariat a été signée entre les deux entités le 4 juillet 2024[31].

[29] https://www.legifrance.gouv.fr/loda/id/LEGITEXT000006068930/2020-10-15/
[30] https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000801164
[31] Convention cadre de partenariat, 4 juillet 2024

VIGINUM a vu ses missions s'élargir en 2026 avec le décret n° 2026-70 du 11 février 2026[32], qui consacre notamment son rôle en matière de détection, de caractérisation et de documentation des modes opératoires informationnels participant aux risques systémiques identifiés dans le cadre du RSN. Ces missions permettent de soutenir et d'appuyer l'Arcom en sa qualité de coordinateur pour les services numériques en France.

Contribution de VIGINUM

Depuis une dizaine d'années, la communauté dédiée à la lutte contre la manipulation de l'information (LMI) se structure, et ses membres (entités gouvernementales, régulateurs, médias, ONG, etc.) multiplient les initiatives pour adopter des concepts communs permettant de mieux comprendre, analyser, et décrire la menace informationnelle numérique.

Si certains de ces concepts sont déjà bien appréhendés et exploités par l'écosystème, comme ceux d'« opération informationnelle » ou de « tactiques, techniques & procédures » (TTPs), les experts décrivent néanmoins de manière différente les ensembles d'éléments techniques, comportementaux et contextuels qu'ils observent – alors même qu'ils disposent de capteurs et de méthodologies d'analyse similaires. Cette absence de grammaire opérationnelle commune est susceptible de nuire à la bonne compréhension de la menace informationnelle.

Afin de résoudre ces difficultés et éventuelles confusions, VIGINUM a élaboré le concept de **« mode opératoire informationnel » (MOI)**, et son équivalent en langue anglaise, « Information Manipulation Set » (IMS). VIGINUM définit un MOI comme un **ensemble de comportements, d'outils et de tactiques, techniques et procédures (TTPs) adverses présumés liés au même acteur malveillant ou groupe d'acteurs malveillants, qui peut être inconnu**.

VIGINUM propose que le concept de MOI soit adopté et employé par la communauté de la LMI, et pour en faciliter l'identification parmi les ensembles qu'ils suivent, invite ses membres à les confronter à deux conditions cumulatives qui constituent l'épine dorsale du concept : la clandestinité et la coordination.

La première condition, dite de « clandestinité », insiste sur les intentions malveillantes des opérateurs : un ensemble peut être considéré comme un MOI si et seulement si les opérateurs engagent des efforts manifestes pour éviter l'imputation, c'est-à-dire le rattachement des infrastructures numériques entre elles et leur attribution à un acteur malveillant. La clandestinité s'opère donc entre l'acteur malveillant et les infrastructures numériques sous leur contrôle, et permet d'exclure du concept tous les ensembles liés à des acteurs malveillants agissant en leur nom propre, ainsi que les ensembles non attribués disposant de différentes infrastructures numériques avec la même bannière.

La seconde condition, dite de « coordination », insiste quant à elle sur le degré d'interaction entre les infrastructures numériques employées : un ensemble peut être considéré comme un MOI si et seulement si les infrastructures numériques sont employées de manière coordonnée, et présumées opérées par le même acteur ou groupe d'acteurs travaillant pour le même commanditaire. Cette condition permet d'assurer la cohérence des éléments techniques, qui doivent posséder des liens techniques ou comportementaux forts.

[32] https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000053458618

Le concept de MOI répond avant tout à un besoin de clarification et a pour principaux avantages :

- d'offrir une dénomination simple, adaptée et centrée sur les éléments techniques, comportementaux et contextuels identifiés et suivis par la communauté ;
- de fonctionner même si les attaquants ne sont pas connus, ni présager de leur origine, de leurs intentions ou de leur niveau de ressources (humaines, financières et techniques) ;
- d'être interopérable avec des concepts éprouvés issus de la *Cyber Threat Intelligence*, ainsi que les standards et les outils déjà exploités au sein de la communauté de la LMI, comme le modèle STIX et *OpenCTI*.

Pour davantage d'informations, le rapport de VIGINUM sur le concept de MOI est accessible au lien suivant : https://www.sgdsn.gouv.fr/publications/definitions-et-objectifs-du-concept-de-mode-operatoire-informationnel-moi

PARTIE 2. LES RISQUES SYSTÉMIQUES DE MANIPULATION DE L'INFORMATION SUR LES GRANDES PLATEFORMES

En application de l'article 34 du RSN, les grandes plateformes sont tenues de recenser, analyser et évaluer « *tout risque systémique au sein de l'Union découlant de la conception ou du fonctionnement de leurs services et de leurs systèmes connexes, y compris des systèmes algorithmiques, ou de l'utilisation faite de leurs services* ». Cette évaluation est réalisée au moins une fois par an et, en tout état de cause, avant de déployer des fonctionnalités susceptibles d'avoir une incidence critique sur les risques systémiques recensés par l'article susmentionné. Elle donne lieu à la publication d'un rapport exposant les résultats de l'évaluation des risques et les mesures mises en place pour les atténuer (article 42 du RSN).

Pour 2025, les grandes plateformes mentionnent dans leur évaluation que plusieurs tendances, similaires à celles observées en 2024, continuent d'avoir des impacts sur les risques systémiques : un nombre important d'élections dans l'Union européenne, des conflits régionaux, l'utilisation croissante de l'intelligence artificielle générative et des opérations d'ingérences étrangères persistantes.

Par ailleurs, ils identifient des facteurs d'influence de ces risques, tels que listés à l'article 34 du RSN, qui renvoient aux enjeux listés dans le tableau suivant :

Facteur d'influence	Manifestations du facteur d'influence Enjeux identifiés par les grandes plateformes
Système de recommandation	- impact de la conception des systèmes sur la priorisation du contenu que les utilisateurs voient - traitement des recommandations émergentes - biais potentiels des algorithmes
Système de modération des contenus	- impact sur la quantité et le type de contenu disponible pour les utilisateurs - impact sur la liberté d'information et le pluralisme
Conditions générales d'utilisation (CGU)	- impact des CGU sur le comportement des utilisateurs - impact de l'application plus ou moins effective des CGU
Systèmes publicitaires	- impact des systèmes de sélection et de présentation de la publicité sur l'exposition des utilisateurs, leur vie privée et la protection des consommateurs - impact sur le modèle économique d'acteurs de la désinformation à des fins purement lucratives - contenus publicitaires déguisés en contenus organiques
Pratiques liées aux données	- usurpations d'identité - impacts sur la protection des consommateurs contre la fraude

Manipulation intentionnelle du service	- création de faux comptes - exploitation des systèmes pour échapper à la détection (contournement) - comportements inauthentiques coordonnés sur le service - liens de *phishing* vers des logiciels malveillants et spam
Intelligence artificielle générative	- risque d'utilisation abusive ou d'exploitation par des tiers - production et diffusion de *deepfakes* - volumes de contenus diffusés rapidement et à grande échelle
Considérations linguistiques et régionales	- enjeu de modération du contenu dans toutes les langues de l'UE - biais linguistiques (traduction)

Néanmoins, la déclinaison des risques systémiques et de leurs facteurs d'influence diffère d'une grande plateforme à l'autre, ces dernières répondant à des problématiques diverses, qui dépendent notamment de leur nature (réseau social, moteur de recherche ou encore plateforme de partage de vidéos). Une analyse par grande plateforme permet ainsi d'avoir un aperçu de leur propre évaluation des risques, qui prend en compte leur écosystème unique.

En lien avec la lutte contre la manipulation de l'information, les grandes plateformes étudiées recensent principalement deux risques systémiques dans leurs rapports[33], tirés de l'article 34 du RSN :

- un risque lié à « *tout effet négatif réel ou prévisible sur le discours civique, les processus électoraux et la sécurité publique* » ;
- un risque lié à « *tout effet négatif réel ou prévisible pour l'exercice des droits fondamentaux, en particulier […] le droit à la liberté d'expression et d'information* ».

Ces deux risques illustrent l'équilibre que le RSN entend préserver entre la lutte contre la manipulation de l'information et les atteintes à la liberté d'expression. À cet égard, la mise en œuvre de mesures d'atténuation des risques ne peut se traduire par des restrictions disproportionnées (ex. modération excessive de contenus licites) qui seraient susceptibles de porter atteinte à l'exercice de cette liberté (cf. article 35 du RSN cité *supra*).

Le règlement ne prévoit pas de méthode harmonisée d'évaluation des risques systémiques ni de lignes directrices en la matière. Les analyses de risques publiées par les grandes plateformes ne sont pas harmonisées à date. Si les évaluations des grandes plateformes reposent sur des approches en partie comparables permettant de calculer

[33] Les rapports de 2024 et 2025 ont été pris en compte pour l'analyse.

des niveaux de risques « *inhérents* »[34] et « *résiduels* »[35], ces risques sont mesurés[36], classés[37] et exprimés[38] de manière différente d'un rapport à l'autre. Ainsi, les niveaux de risques identifiés ne sont pas comparables entre plateformes.

Enfin, certaines grandes plateformes présentent leur approche d'évaluation des risques sans entrer dans le détail des méthodes de calcul (Google, Meta). Quant à Wikipédia, seule une description des risques est fournie, sans évaluation quantifiée.

Avertissement

L'évaluation des risques systémiques identifiés par les grandes plateformes sur leurs services relève de leur responsabilité. La présentation proposée dans ce bilan ne saurait être interprétée comme une validation ou une prise de position de l'Arcom sur la manière dont les plateformes ont choisi de mener leurs analyses de risques.

La supervision du respect par les grandes plateformes de leur obligation d'évaluer et d'atténuer les risques systémiques a été initiée fin 2024 avec la publication de leurs premiers rapports. Elle relève en premier lieu de la Commission européenne qui seule peut sanctionner des manquements. Mais cette supervision associe tous les coordinateurs pour les services numériques (notamment via le comité européen des services numériques ou « *DSA board* »), de multiples autorités publiques, le monde académique et la société civile.

Elle doit monter en puissance à chaque itération annuelle, avec la mobilisation croissante d'informations qui découlent des obligations de transparence imposées aux plateformes. Ainsi, le droit d'accès des chercheurs à des données non-publiques pour procéder à des contre-évaluations des travaux des grandes plateformes sur les risques systémiques ne commencera à être opérationnel qu'en 2026.

[34] Le risque inhérent au service désigne le niveau de risque intrinsèque associé à l'utilisation, au fonctionnement ou à la conception d'un service avant toute mesure d'atténuation. Autrement dit, il s'agit du risque qui découle naturellement de ce que le service permet, expose ou rend possible, indépendamment des contrôles, protections ou politiques internes qui pourraient ensuite le réduire.
[35] Le risque résiduel au service désigne le niveau de risque qui subsiste après la mise en place de toutes les mesures d'atténuation. Il correspond donc à la part de risque qui ne peut être entièrement éliminée malgré les mesures mises en œuvre.
[36] Par exemple, Snapchat retient des critères pour calculer la gravité du risque (« *Scope* » ; « *Scale* » ; « *Remediability* »), qui sont différents de ceux de Meta (« *Scale* » ; « *Harm Type* » ; « *Vulnerability of impacted people* » ; « *Intentionality* ») et Google (« *Scope* » et « *Impact* »).
[37] Par exemple, certaines grandes plateformes procèdent à la création de catégories différant légèrement des risques tels qu'ils sont décrits à l'article 34 du RSN. TikTok inclut la sécurité publique dans la catégorie « *harmful misinformation* » alors que les autres grandes plateformes l'ont intégrée avec les discours civiques et les processus électoraux (comme le mentionne le RSN). Snapchat ne prend pas en compte les mêmes catégories que les autres grandes plateformes pour les risques systémiques relatifs au discours civique et aux processus électoraux (« *les effets négatifs sur la démocratie et les processus électoraux* » sont distingués des « *effets négatifs sur le discours civique* »). En outre, certaines grandes plateformes subdivisent les risques systémiques prévus par le RSN en « *sous-risques* » : Meta décline ainsi les risques systémiques en sous-risques (« *problem area* »), comme la désinformation, les comportements inauthentiques, le spam, etc.
[38] Par exemple, certaines grandes plateformes proposent une échelle allant de 1 à 5, d'autres de 1 à 3.

I. Les risques pour le discours civique, les processus électoraux et la sécurité publique

L'étude des rapports sur les risques systémiques publiés en 2025 fait ressortir un ensemble de risques liés à cette catégorie, généralement liés :

- à la désinformation électorale ;
- à des détournements de comptes, des tentatives de *phishing* ;
- à la manipulation coordonnée, aux activités inauthentiques et aux ingérences étrangères ;
- aux biais algorithmiques, entrainant des effets de polarisation politique ;
- à la diffusion de contenus synthétiques trompeurs, de *deepfakes* ;
- au ciblage de publics vulnérables ;
- à l'intimidation et au harcèlement ;
- au piratage et à la fuite de données.

L'ensemble des grandes plateformes étudiées distingue les risques pour le « *discours civique et les processus électoraux* » des risques identifiés en matière de « *sécurité publique* » :

- **pour les risques portant sur le discours civique et les processus électoraux**, la majorité des grandes plateformes identifie :
  - un risque inhérent oscillant entre élevé et moyen ;
  - un risque résiduel faible, exception faite de YouTube qui considère que le niveau de risque résiduel reste élevé, « *malgré son état de préparation, en grande partie en raison de facteurs externes, notamment la nature dynamique et virale des informations trompeuses dans le domaine politique, pendant les élections et en période de crise et de troubles civils* »[39] ;
- **pour les risques en matière de sécurité publique**, le risque inhérent est très variable d'une plateforme à l'autre, alors que le niveau de risque résiduel oscille entre moyen et faible.

Sans pouvoir entrer dans une démarche comparative fiable (eu égard à la diversité des méthodes d'évaluation et de calcul employées par les grandes plateformes), l'Arcom dresse le constat que chacun d'entre eux estime que les mesures d'atténuation mises en place sur leur service ont permis de réduire les niveaux de risques identifiés à un niveau plutôt faible (sauf YouTube). En revanche, l'homogénéité des niveaux de risque résiduel, qui apparaissent à un niveau similaire sur la plupart des services, peut interroger dans un contexte où les caractéristiques des plateformes – en termes de modèle économique, d'architecture technique, de fonctionnalités, de publics visés ou encore de volumes et de nature des contenus diffusés – diffèrent sensiblement.

L'approche de Meta

Dans le cadre de ses évaluations des risques réalisées pour les catégories « *contenus trompeurs* » et « *discours civique et processus électoraux* », Meta souligne que l'essor des nouvelles technologies, en particulier l'IA générative, offre aux acteurs malveillants de nouveaux moyens de cibler certains publics à l'aide de *TTPs* en constante évolution. Meta observe également que les opérations d'ingérence étrangère et les comportements inauthentiques coordonnés demeurent persistants sur ses services, s'accompagnant

[39] Extrait du rapport sur les risques systémiques de Google, publié en décembre 2025.

d'abus des systèmes de sécurité et de signalement de Facebook et Instagram ainsi que de tentatives de contournement des dispositifs de détection.

L'écosystème attaché aux comportements inauthentiques évoluant rapidement, Meta considère que leur identification devient plus complexe, certaines campagnes d'influence se déplaçant même vers d'autres espaces en ligne. Le fournisseur de services note par ailleurs que le nombre élevé d'élections au sein de l'Union européenne pourrait amplifier les risques de comportements inauthentiques, touchant notamment des partis politiques, y compris via l'utilisation de l'IA générative.

Meta relève enfin une intensification des actions visant à tromper les utilisateurs par la fraude, la désinformation ou les *deepfakes*.

L'approche de X

Selon la plateforme, plusieurs facteurs créent un environnement de risque propice à la diffusion de contenus nuisibles et à l'ingérence dans les processus démocratiques. Les élections dans l'Union européenne et divers conflits régionaux peuvent inciter des acteurs malveillants à exploiter la plateforme pour diffuser des informations fausses ou trompeuses, notamment sur les élections, ou pour mener des attaques coordonnées menaçant la sécurité publique.

Ce contexte est aggravé par la possibilité de formation de chambres d'écho, où les utilisateurs sont principalement exposés à des contenus confirmant leurs croyances préexistantes, renforçant ainsi leurs biais cognitifs.

Par ailleurs, les systèmes de recommandation de X – y compris ceux offrant une amplification accrue via l'abonnement ou la publicité – peuvent involontairement accroître la visibilité de contenus trompeurs.

Pour réduire ce risque systémique, X met en avant la méthode des Notes de la Communauté, rédigées par des usagers volontaires, afin de contextualiser les contenus qui apparaissent trompeurs aux usagers eux-mêmes. X n'évalue pas, à ce jour, l'efficacité de cette démarche dans les différentes langues de l'Union européenne.

Enfin, X reconnaît que les risques pesant sur le discours civique et les processus démocratiques peuvent résulter d'autres menaces, comme l'intimidation des électeurs, les discours haineux, les ingérences étrangères et les comportements inauthentiques.

X précise néanmoins que la relation entre les contenus nuisibles sur la plateforme et les actions hors ligne est complexe et la causalité est difficile à déterminer.

L'approche de Google[40] (pour Google Search et YouTube)

Google Search et YouTube n'ayant pas la même nature – l'un étant un moteur de recherche et l'autre une plateforme de partage de vidéos – les risques listés par Google dans ses rapports sont très divers.

[40] À noter que les rapports sur les risques systémiques de Google prennent en compte tous ses services (à la différence de Microsoft et de Meta qui ont un rapport dédié pour chacun de leurs services). Google discrimine donc peu les risques identifiés par services, bien que certains soient propres à Google Search.

Sur ces services, les informations fausses ou trompeuses peuvent prendre des formes variées : pages conçues pour diffuser ou monétiser du contenu trompeur via les services publicitaires de Google, vidéos manipulées sur YouTube ou encore sites web diffusant du contenu de faible qualité et apparaissant dans les résultats de recherche. Ces pratiques s'inscrivent dans un éventail plus large de comportements abusifs, qui comprend également la fraude, l'usurpation d'identité ou l'utilisation de l'IA à des fins malveillantes (*deepfakes*).

Google observe en particulier une multiplication d'escroqueries plus sophistiquées, portées par des acteurs qui opèrent à grande échelle, adaptent continuellement leurs méthodes et combinent approches en ligne et hors ligne. Si l'accessibilité accrue des outils d'IA n'a pas créé de nouveaux types de fraudes, elle a renforcé l'efficacité de vecteurs existants : les modèles linguistiques avancés ainsi que les générateurs d'images, de vidéos et d'audios sont désormais également utilisés pour produire des contenus frauduleux plus convaincants et plus difficiles à détecter.

Sur Google Search, les risques incluent également la présence potentielle de désinformation électorale ou relative à d'éventuels troubles sociaux, ainsi que des menaces numériques telles que le détournement de compte, le *phishing* ou des campagnes d'influence visant les utilisateurs à des moments clés du débat démocratique. S'y ajoutent les risques liés à un mauvais usage ou à des défaillances de grands modèles d'apprentissage automatique, à la persistance de contenus illégaux ou nuisibles et à la présence d'informations de faible qualité.

Pour YouTube, Google identifie aussi des risques liés au comportement des utilisateurs, qui pourraient mettre en ligne du contenu violent, dangereux, sensible ou trompeur.

L'approche de TikTok

Contrairement à d'autres services qui insistent davantage sur les *TTPs* utilisées pour manipuler l'information, TikTok se concentre en particulier sur les types de contenus de désinformation susceptibles d'affecter le discours civique et les processus électoraux[41]. La plateforme prend par exemple en compte les informations erronées sur les conditions de vote ou sur la date du scrutin. TikTok identifie également des contenus avançant de fausses affirmations au sujet des conditions d'éligibilité des candidats ou de responsables déjà élus, des allégations de fraude électorale (comme la manipulation des machines à voter) ou des contenus prétendant qu'une élection a été ou sera truquée, sapant ainsi la confiance des électeurs dans les résultats.

À cela s'ajoutent les théories du complot visant des candidats, des tentatives d'usurpation d'identité de personnalités politiques, ainsi que l'utilisation de médias synthétiques ou manipulés mettant en scène des figures politiques, susceptibles d'influencer l'opinion publique.

TikTok relève que ces risques peuvent être amplifiés par divers phénomènes : opérations d'influence, piratages et fuites de données, publicité frauduleuse, faux engagements, ou encore spam. La plateforme mentionne que le contexte géopolitique peut également exacerber ces risques.

[41] À noter que toutes les grandes plateformes prennent en compte la désinformation électorale dans leurs CGU, bien que celles-ci n'en fassent pas autant état que TikTok dans leur rapport sur les risques systémiques.

L'approche de Microsoft Bing

Bien que les utilisateurs de Microsoft Bing[42] ne puissent pas publier ou partager des informations sur le moteur de recherche (ce qui minimise le risque que le contenu devienne viral ou entraine des préjudices à grande échelle concernant le discours civique, les processus électoraux et la sécurité publique), Bing mentionne une série de risques relatifs à la nature de son service, notamment :

- que le contenu indexé par Bing et qui provient de tiers sur le web inclut du contenu nuisible. Autrement dit, si ce risque n'est pas suffisamment atténué, ses utilisateurs peuvent être exposés à du contenu de faible qualité ou trompeur (ex. de la désinformation électorale) dans les résultats de recherche ;
- que le moteur de recherche reproduise du contenu de faible autorité, de mauvaise qualité, obsolète ou incluant des informations inexactes et trompeuses sur les processus électoraux, les procédures de vote ou les résultats d'un scrutin ;
- que les *data voids* permettent la mise en avant de contenu de faible qualité ;
- que les systèmes de sécurité de Bing introduisent des biais ou limitent l'accès aux informations électorales importantes ;
- que les résultats de recherche favorisent certaines opinions politiques, influencent les perceptions et polarisent davantage les utilisateurs, etc.

Par ailleurs, les fonctionnalités d'IA générative proposées par Microsoft Bing[43] pourraient être utilisées pour consommer ou créer des médias synthétiques, introduire de manière involontaire un biais dans les résultats de recherche ou contribuer à un phénomène de polarisation politique.

Le moteur de recherche mentionne enfin les risques liés à « *l'intimidation, au harcèlement, au ciblage coordonné ou à la coercition à l'encontre d'acteurs politiques ou civiques* », qui peuvent découler des résultats de recherche proposés par son service.

L'approche de LinkedIn

En raison de la nature professionnelle du réseau social, LinkedIn considère que la plateforme est moins exposée que les autres grandes plateformes aux contenus susceptibles d'avoir un impact négatif sur le discours civique et les processus électoraux. Néanmoins, elle identifie certaines menaces qu'elle inclut dans ce risque systémique : la désinformation électorale, la manipulation coordonnée, les activités inauthentiques, la perturbation des processus électoraux, les ingérences étrangères, la création de chambres d'écho qui peuvent polariser les utilisateurs, le ciblage de groupes vulnérables ou encore les biais liés aux systèmes de recommandation.

L'approche de Snapchat

Si Snapchat relève une faible prévalence des contenus de désinformation et de *deepfakes* sur le service (même si les fonctionnalités « *Spotlight* »[44] et « *Lens For You* »[45] peuvent en contenir), la plateforme mentionne que les risques liés au discours civique, aux processus électoraux et à la sécurité publique peuvent inclure une

[42] Et de tous les moteurs de recherche de façon générale.
[43] Copilot, Bing Image and Video creator.
[44] *Spotlight* est un espace du service proposant des vidéos de courte durée.
[45] *Lens For You* est une fonctionnalité de recommandation personnalisée de Lenses (filtres de réalité augmentée).

polarisation des discours en ligne. À ce titre, le potentiel de contenu personnalisé et les biais algorithmiques peuvent enfermer les utilisateurs dans des bulles informationnelles (qui comportent notamment des contenus extrêmes ou sensationnalistes visant à retenir leur attention).

Concernant la sécurité publique, Snapchat considère que les contenus dangereux, préjudiciables et incitatifs pourraient contribuer à amplifier le risque s'ils sont largement diffusés. La plateforme mentionne une liste de contenus pouvant affecter la sécurité publique, comme les discours de haine, les contenus de mésinformation préjudiciable, les contenus encourageant à la violence ou à un comportement dangereux.

L'approche de Wikipédia

Wikipédia identifie plusieurs risques spécifiques, étroitement liés à son modèle ouvert et collaboratif. Concernant le discours civique et les processus électoraux, la plateforme mentionne que des acteurs cherchant à influencer un résultat politique pourraient mener des campagnes coordonnées visant à insérer du contenu trompeur dans les articles de l'encyclopédie. Un tel phénomène – déjà observé, notamment sur la version croate de Wikipédia en 2021[46] – peut réduire sa fiabilité globale, induire les lecteurs en erreur et contribuer à la propagation de récits manipulés. Ce cas peut également viser des contenus historiques et géographiques, qui peuvent être modifiés pour servir certaines idéologies.

Par ailleurs, Wikipédia souligne un risque lié à l'usage croissant de l'IA générative et des technologies d'apprentissage automatique par les contributeurs bénévoles. Si ces outils peuvent faciliter la rédaction et la traduction d'articles, ils sont également susceptibles d'introduire de la mésinformation, notamment en raison de phénomènes « *d'hallucinations* ». Ils peuvent aussi renforcer des biais existants (ex. biais sexistes, ethniques, linguistiques, etc.) et contribuer à la diffusion involontaire de représentations partielles ou discriminatoires.

II. Les risques pour la liberté d'expression et d'information

A. L'évaluation réalisée par les grandes plateformes

Régulation et liberté d'expression

Comme le mentionne la Stratégie nationale de lutte contre les manipulations de l'information 2026-2030[47], certains acteurs peuvent se prévaloir d'une interprétation juridique du droit européen pour supposer un risque d'atteinte à la liberté d'expression et d'information et orienter ainsi les perceptions des utilisateurs sur les objectifs de la régulation européenne. Ces normes consacrent pourtant pleinement ces libertés en ligne, et leur application n'a pas pour objet de les restreindre, mais au contraire de les protéger dans l'environnement numérique.

Les grandes plateformes identifient un certain nombre de risques pour la liberté d'expression et d'information, qui correspondent notamment à la possibilité que leurs systèmes de modération, leurs algorithmes, leurs politiques internes ou d'influences

[46] https://meta.wikimedia.org/wiki/Croatian_Wikipedia_Disinformation_Assessment-2021
[47] V. note 6.

extérieures, portent atteinte au droit fondamental des utilisateurs à s'exprimer et s'informer librement.

Ils identifient notamment les risques suivants :

- la sur-modération ou la suppression injustifiée de contenus légitimes ;
- la restriction excessive de l'accès à l'information ;
- les atteintes au pluralisme ;
- les biais algorithmiques ;
- la manipulation des algorithmes par des acteurs malveillants.

Ces risques englobent également des phénomènes tels que l'atrophie du débat public (autocensure, découragement des citoyens, réduction de la participation), la désinformation ou les campagnes coordonnées visant à orienter ou limiter l'expression en ligne.

Ils sont intimement liés aux risques relatifs au discours civique, aux processus électoraux et à la sécurité publique, car la qualité du débat démocratique dépend directement de la capacité des citoyens à accéder à une information fiable et pluraliste et à exprimer librement leurs opinions. Toute limitation injustifiée à ces libertés fondamentales pourrait porter atteinte à la participation éclairée des citoyens dans la vie démocratique ainsi qu'à l'intégrité des scrutins.

La majorité des grandes plateformes étudiées identifie, concernant la liberté d'expression et d'information, des niveaux de :
- « *risque inhérent* » considérés comme élevés ;
- « *risque résiduel* » considérés comme faible.

L'Arcom constate, comme pour les risques liés pour le discours civique, les processus électoraux et la sécurité publique (*supra*), que chacun d'entre eux considère que les mesures d'atténuation mises en œuvre pour réduire ce risque ont été particulièrement efficaces.

L'approche de Meta

Pour ses services, Meta considère que la modération excessive des contenus légitimes constitue une préoccupation majeure, en ce qu'elle restreint de manière injustifiée la liberté d'expression des utilisateurs. Cette situation pourrait être accentuée par des limitations linguistiques, qui rendent la modération des contenus plus complexe.

Facebook et Instagram pourraient aussi être confrontées à des actions pouvant décourager l'expression, telles que des comportements qui limitent la participation des utilisateurs ou les dissuadent de s'exprimer librement. Meta rappelle que l'équilibre entre la liberté d'expression et la sécurité des utilisateurs représente un défi important, car il s'agit avant tout de protéger les citoyens tout en leur garantissant l'accès à l'information.

Par ailleurs, il est à noter que des acteurs malveillants exploitent régulièrement les fonctionnalités de Meta pour empiéter sur la liberté d'expression, par le biais de spam, de signalement abusif, de harcèlement ou d'intimidation, ce qui peut perturber le débat public et restreindre l'expression des utilisateurs.

Meta doit aussi maintenir la cohérence de ses politiques internes avec les cadres juridiques nationaux en vigueur ainsi que les normes culturelles des États, et répondre aux injonctions de retrait gouvernementales.

Enfin, certains facteurs contextuels (similaires à ceux identifiés en 2024), comme le nombre élevé d'élections dans l'Union européenne et les conflits régionaux, pourraient accroitre la pression sur la liberté d'expression, même si Meta a évalué que ces éléments n'ont pas modifié le niveau de risque inhérent pour 2025.

À noter que cette année, Meta a particulièrement constaté un changement dans l'environnement du risque, dû à une augmentation de la sur-application des lois et des faux positifs causés par les systèmes de plus en plus complexes nécessaires à la modération du contenu.

L'approche de X

X relève plusieurs menaces qui s'inscrivent directement dans le champ des risques systémiques relatifs à la liberté d'expression et d'information. Selon la plateforme, la manipulation de l'information (qu'il s'agisse d'opérations d'influence, de diffusion de récits trompeurs ou encore d'amplification de certaines voix) constitue un facteur majeur pouvant affecter le pluralisme des informations accessibles aux utilisateurs. Ces risques peuvent être accentués par des actions menées en dehors de la plateforme, telles que des stratégies de coordination destinées à simuler artificiellement un engagement ou à influencer des tendances.

Par ailleurs, des campagnes de signalements massifs et malveillants, souvent conçues pour provoquer une application disproportionnée de la modération, peuvent entraîner la suppression de contenus légitimes ou la suspension injustifiée de comptes.

L'ensemble de ces phénomènes compromet la fiabilité de l'espace informationnel offert par X et la capacité des utilisateurs à participer à un débat public ouvert, authentique et pluraliste, illustrant concrètement l'impact potentiel sur la liberté d'expression et d'information.

L'approche de Google (pour Google Search et YouTube)

Sans être discriminés, les risques liés aux services de Google, qui peuvent impacter la liberté d'expression et d'information, sont les suivants :

- la suppression de contenus ayant une valeur probante, par exemple dans le cadre d'une procédure judiciaire ;
- la suppression disproportionnée ou non nécessaire de contenus ;
- le fait, pour les utilisateurs, de rencontrer des obstacles dans le processus de signalement de contenus potentiellement illicites ou dans le processus visant à contester la suppression d'un contenu, ce qui limite leur capacité à participer pleinement au contrôle de l'information disponible ;
- un manque de transparence ou d'options pouvant compromettre la capacité des utilisateurs à prendre des décisions autonomes et éclairées quant aux contenus qu'ils consultent.

Google note par ailleurs que la visibilité des contenus sur ses services peut avoir un effet direct sur le pluralisme des médias, en influençant la diversité, la polarisation et la pluralité des points de vue auxquels sont exposés les utilisateurs.

L'approche de TikTok

TikTok est l'une des seules plateformes à mentionner peu d'éléments sur les risques liés à la liberté d'expression et d'information en particulier, bien que celle-ci soit évoquée plus généralement dans les risques d'atteinte aux droits fondamentaux.

La plateforme précise notamment qu'il y a un risque que les utilisateurs comprennent mal ou ignorent ses CGU et adoptent des comportements interdits sur la plateforme. Elle mentionne aussi les risques liés à limitation de la visibilité de certaines publications, ainsi qu'à la sur-modération et à la sous-modération des contenus.

L'approche de Microsoft Bing

Microsoft identifie plusieurs risques susceptibles d'affecter la liberté d'expression et d'information sur Bing, qui ne sont pas propres à la nature du service :

- la sur-modération ou le sur-blocage de contenus légitimes ;
- le risque que ses systèmes de classement favorisent ou au contraire rétrogradent de manière disproportionnée certains types de contenus ;
- le blocage de contenus en réponse aux injonctions de retrait gouvernementales et aux demandes des forces de l'ordre ;
- la restriction disproportionnée de l'accès à l'information ;
- une atteinte au pluralisme ;
- une atrophie du débat public.

Si Bing estime que la probabilité inhérente de risque systémique pour la liberté d'expression et d'information demeure « *lointaine* », elle est toutefois plus élevée en ce qui concerne le risque de restriction excessive via les fonctionnalités reposant sur l'IA générative. Selon le moteur de recherche, ce risque serait néanmoins atténué par la possibilité, pour les utilisateurs, de continuer à accéder à l'information via la recherche web ou d'autres services.

L'approche de LinkedIn

LinkedIn identifie plusieurs risques liés à la liberté d'expression et d'information qui peuvent se manifester en l'absence de mesures d'atténuation suffisantes : la suppression de contenus légitimes ou, à l'inverse, l'amplification disproportionnée de certaines voix, qui pourraient altérer le pluralisme de l'information et la diversité des opinions.

LinkedIn considère la gravité inhérente à ces risques comme étant élevée, compte tenu des dommages possibles pour les systèmes sociaux, le bien-être collectif, les processus politiques et la capacité des citoyens à s'informer librement.

Pour la dernière période d'évaluation des risques systémiques[48], les données des recours suggèrent toutefois une certaine stabilité du système de modération de LinkedIn : seuls 4,6 % des décisions de modération ont été contestés par les utilisateurs, et LinkedIn a rétabli environ 22,63 % des contenus concernés, soit 1,49 % de l'ensemble des décisions de modération[49].

[48] Août 2024 – Août 2025.
[49] LinkedIn est la seule plateforme à fournir des données chiffrées à ce sujet.

L'approche de Snapchat

Snapchat considère que ses systèmes peuvent être manipulés de deux manières principales :

- les utilisateurs pourraient chercher à partager du contenu illégal et/ou contraire à ses CGU, qui échapperait au système de détection ;
- les utilisateurs pourraient abuser de son système de modération et signaler de mauvaise foi des comptes ou des contenus non violents.

L'approche de Wikipédia

Wikipédia considère que les risques identifiés pour le discours civique et les processus électoraux sont similaires à ceux qui pourraient porter atteinte à la liberté d'expression et d'information.

B. L'apport du RSN en matière de protection de la liberté d'expression et d'information

Le RSN vise à concilier la lutte contre les contenus illicites et préjudiciables avec la protection des libertés fondamentales des utilisateurs en ligne, comme la liberté d'expression et d'information.

À ce titre, le règlement oblige les grandes plateformes à :

- **informer les utilisateurs des décisions qu'elles prennent en matière de modération des contenus et à expliquer les motifs de ces décisions** (art.17 du RSN). Les plateformes en ligne doivent intégrer ces exposés des motifs dans une base de données centralisée qui est accessible en ligne[50], accompagnés des informations fournies à l'utilisateur conformément à l'article 17 du RSN ;

- **donner l'accès, aux destinataires du service, à un système de recours interne** leur permettant de demander le réexamen d'une décision prise par la plateforme, qu'il s'agisse d'une restriction qu'elle a opérée sur un contenu ou le compte de l'utilisateur ou des suites qu'elle a données à un signalement qu'il a effectué (article 20 du RSN). Le fournisseur de la plateforme doit informer le plaignant dans les meilleurs délais de sa décision à l'issue du réexamen, de façon motivée, et en s'assurant qu'elle a été prise « *sous le contrôle de collaborateurs dûment qualifiés et pas uniquement par des moyens automatisés* » ;

- **mettre à la disposition du public, tous les six mois, « *des rapports clairs et facilement compréhensibles sur les éventuelles activités de modération des contenus auxquelles ils se sont livrés* »** en réponse à des injonctions et des signalements ou dans le cadre de mesures proactives, ainsi que sur leurs relations avec les organes de règlement extra-judiciaires des litiges (OREL) visés à l'article 21 du RSN, qui permettent d'offrir aux utilisateurs des garanties supplémentaires visant à renforcer la protection de leurs droits fondamentaux en ligne.

[50] https://transparency.dsa.ec.europa.eu/statement?lang=fr

La montée en puissance du RSN amène les grandes plateformes à continuer à renforcer la transparence sur leurs outils de modération pour favoriser leur meilleure compréhension. Les premiers rapports de transparence harmonisés attendus en 2026 devraient permettre de documenter de manière plus homogène les pratiques de modération et d'apprécier plus finement l'équilibre effectif entre les mesures prises à l'encontre des contenus signalés et le respect de la liberté d'expression.

La Commission européenne, en association avec les coordinateurs pour les services numériques, se saisit des analyses de risques systémiques et des rapports de transparence des grandes plateformes pour contrôler leur conformité au RSN dans le cadre de procédures qui, en cas de suspicion de manquement, peuvent les amener à faire évoluer et adapter leurs services. Les perspectives encourageantes de ce cercle vertueux de régulation sont sans préjudice des défis à venir dans un contexte où certains outils prévus par le RSN, comme l'accès aux données de l'article 40 du règlement, sont en cours de déploiement afin de permettre un suivi efficace de l'ampleur des risques d'effets négatifs sur le discours civique sur les plateformes.

***

Dans l'ensemble, les risques systémiques liés au discours civique, aux processus électoraux, à la sécurité publique et à la liberté d'expression et d'information apparaissent véritablement interconnectés. Les menaces identifiées dans ces deux catégories se renforcent mutuellement et influencent directement la qualité du débat public. Les différentes plateformes étudiées en sont conscientes et intègrent ces enjeux dans leur évaluation et leurs CGU, même si les approches demeurent encore hétérogènes et façonnées par la nature des services. Malgré cette absence d'uniformité, se dégage un constat commun, qui repose sur la nécessité d'anticiper, d'atténuer et de surveiller sur le long terme ces risques, afin de préserver un environnement numérique sûr, fiable et compatible avec les droits fondamentaux des utilisateurs.

L'Arcom note toutefois que les facteurs de risques systémiques découlant de la conception et du fonctionnement des systèmes algorithmiques des grandes plateformes sont très peu documentés par ces dernières, alors même qu'elles reposent sur des algorithmes opaques et (ultra)personnalisés, conçus pour maximiser l'engagement des utilisateurs sur leurs services.

Il en est de même en ce qui concerne les facteurs liés aux considérations linguistiques et régionales, qui ne font pas l'objet de développements suffisants. Pourtant, celles-ci pourraient également poser des difficultés aux utilisateurs, comme l'a mentionné AI Forensics dans une étude publiée en août 2025[51] visant YouTube. Ces travaux ont en effet révélé que les fonctionnalités de sécurité introduites par la plateforme (panneaux d'information contextuelle thématique et panneaux d'information fournissant le contexte de l'éditeur) étaient appliquées de manière incohérente dans les différents pays européens, en particulier s'agissant de sujets qui peuvent être soumis à des théories du complot (vaccination, changement climatique, etc.). D'après l'étude, le critère linguistique serait le principal défaut de ces fonctionnalités, les tests d'AI Forensics ayant permis d'affirmer que YouTube accordait « *une attention disproportionnée aux langues occidentales* ».

[51] YouTube's Safety Features Lost in Translation | AI Forensics

PARTIE 3. LES MESURES PRISES PAR LES GRANDES PLATEFORMES POUR PROTÉGER L'INTÉGRITÉ DE LEUR SERVICE

Face à l'ensemble des risques identifiés par les grandes plateformes, ces dernières doivent mettre en place « *des mesures d'atténuation raisonnables, proportionnées et efficaces, adaptées […], en tenant compte en particulier de l'incidence de ces mesures sur les droits fondamentaux* » (article 35 du RSN). Ces mesures peuvent prendre des formes variées, l'article susmentionné en produisant une liste non exhaustive.

Parmi celles-ci, ce bilan aura vocation à évoquer les conditions générales d'utilisation des grandes plateformes dédiées à la lutte contre les *TTPs*, l'apport des systèmes de détection automatisés, l'importance des approches humaines et collaboratives, ainsi que les actions éducatives et de sensibilisation mises en place par les grandes plateformes.

Avertissement

- les exemples de mesures d'atténuation cités dans cette partie sont illustratifs et non exhaustifs ;
- les CGU listées ne mentionnent pas les exceptions aux interdictions qu'elles prévoient (ex. comptes de fans, comptes parodiques, comptes à vocation éducative, etc.).
- les développements ci-après reposent sur des éléments déclaratifs des grandes plateformes, sans appréciation de l'Arcom sur l'effectivité de leur mise en œuvre.

I. Lutter contre les techniques de manipulation via les conditions d'utilisation du service

L'Arcom observe que les conditions générales d'utilisation[52] des grandes plateformes en matière de lutte contre les *TTPs* sont particulièrement fournies et détaillées, témoignant d'un effort réel de cadrage des comportements interdits sur les services. Elle souligne toutefois que ces dispositions demeurent hétérogènes d'une grande plateforme à l'autre, tant au niveau de leur précision que des thématiques abordées.

Cette diversité reflète les différences de modèle, de gouvernance et de maturité des mesures d'atténuation des grandes plateformes, mais peut également nuire à la lisibilité et à la comparabilité des situations couvertes par ces CGU, rendant ainsi complexe l'évaluation globale de la cohérence et de l'efficacité des politiques[53].

A. Les médias synthétiques et manipulés

L'approche de Meta

Meta prend en compte les critères suivants pour identifier les médias synthétiques et manipulés :

- un contenu créé ou retouché numériquement ;

[52] Les développements suivants exposent l'essentiel des CGU concernées. L'annexe 1 du bilan en fournit une présentation détaillée par grande plateforme, nourrie d'exemples variés.
[53] Les parties exposées ci-après reprennent les CGU brutes des grandes plateformes étudiées.

- un contenu susceptible d'induire en erreur.

Meta peut placer une étiquette informative sur le média synthétique ou manipulé, lorsqu'il s'agit d'une image, d'une vidéo ou d'un contenu audio qui semble réaliste, créé ou retouché numériquement et qu'il implique un risque particulièrement élevé de tromper considérablement le public sur un sujet d'intérêt général. Meta peut aller jusqu'à rejeter le contenu s'il s'agit d'une publicité.

L'approche de X

X inclut dans la notion de média manipulé les éléments suivants :

- les contenus non authentiques susceptibles de tromper ou de nuire ;
- les médias hors contexte, pouvant semer la confusion sur des questions d'intérêt public, de compromettre la sécurité publique ou de causer un préjudice grave ;
- les médias substantiellement édités ou post-traités ;
- les médias contenant des informations visuelles ou auditives qui ont été ajoutées, modifiées ou supprimées et qui modifient fondamentalement la compréhension, le sens ou le contexte du média ;
- les médias représentant une personne réelle qui a été fabriquée ou simulée, notamment grâce à l'utilisation d'algorithmes ou d'une intelligence artificielle.

L'approche de YouTube

YouTube distingue les contenus « *manipulés* » des contenus « *attribués à tort* ».

Les contenus « *manipulés* » sont des contenus qui ont été techniquement manipulés ou falsifiés de manière à induire les utilisateurs en erreur et qui peuvent présenter un risque important de préjudice majeur.

Les contenus « *attribués à tort* » sont des contenus présentant un risque de préjudice majeur en prétendant faussement qu'une ancienne vidéo d'un événement passé représente un événement actuel.

L'approche de TikTok

TikTok n'autorise pas le matériel visuel ou audio qui a été édité, assemblé, combiné, d'une manière qui pourrait induire un utilisateur en erreur sur des événements du monde réel. La plateforme peut rendre inéligible au fil « *For You* » tout contenu d'apparence réaliste dont il n'a pas encore été confirmé qu'il est du contenu généré par l'IA ou modifié de manière significative, mais qui présente des questions d'importance publique d'une manière qui pourrait conduire à une mauvaise interprétation ou nuire à des personnalités privées. Par ailleurs, tous les contenus présentant des images, des vidéos ou de l'audio entièrement ou considérablement modifiés ou générés par l'IA doivent être présentés comme tels ou étiquetés.

L'approche de Microsoft (pour Bing et LinkedIn)

Microsoft interdit, pour l'ensemble de ses services, la création et la diffusion de contenus trompeurs générés par IA. Cela inclut les contenus audios, vidéos et les images générés par IA qui falsifient ou modifient de manière trompeuse l'apparence, la voix ou les

actions d'autres personnes[54]. LinkedIn interdit les médias manipulés ou synthétiques trompeurs qui ne divulguent pas clairement la nature fausse ou modifiée du contenu.

L'approche de Snapchat

Pour les contenus organiques, Snapchat mentionne simplement qu'il est interdit de générer des *deepfakes*. S'agissant des contenus publicitaires, la plateforme précise qu'elle interdit les contenus qui comprennent des avatars de synthèse ou des ressemblances visuelles ou vocales avec une personne réelle et qui ont été manipulés à des fins frauduleuses ou fallacieuses (que ce soit par le biais d'IA générative ou d'un montage trompeur).

L'approche de Google Search

Google Search n'autorise ni contenu audio, vidéo ou image manipulé dans le but de tromper, frauder ou d'induire en erreur en présentant une version mensongère d'actions ou d'événements qui ne se sont manifestement pas produits. Ceci inclut tout contenu susceptible d'induire une personne raisonnable en erreur quant à sa compréhension ou son interprétation, et pouvant ainsi causer un préjudice important à des groupes ou des individus, ou compromettre gravement la participation ou la confiance dans les processus civiques ou électoraux. Google Search a d'ailleurs une page dédiée aux mesures mises en place contre les *deepfakes*[55].

L'approche de Wikipédia

Wikipédia n'a pas de CGU sur les médias synthétiques et manipulés formellement rédigées comme les autres grandes plateformes. Cependant, ses règles communautaires interdisent l'insertion de contenus trompeurs, ce qui pourrait implicitement couvrir cette catégorie. Cela peut se confirmer grâce au projet communautaire *WikiProject AI Cleanup*[56], qui vise à lutter contre les contenus générés par IA qui sont sans source et de mauvaise qualité. En août 2025, Wikipédia a d'ailleurs adopté une politique permettant aux contributeurs de proposer la suppression rapide des articles suspectés d'être générés par IA.

B. Fraude et usurpation d'identité

1. Fraude

La fraude est étroitement liée aux *TTPs* présentées *supra*, car elle constitue un vecteur par lequel des acteurs malveillants peuvent tromper, manipuler ou exploiter les utilisateurs pour influencer leur perception de l'information.

Toutes les grandes plateformes étudiées interdisent la fraude sur leurs services. Parmi eux :

[54] À noter que Microsoft intègre désormais Copilot au sein de l'ensemble de ses produits, dont Microsoft Bing. Les conditions d'utilisation de Copilot, qui reprennent les politiques générales de Microsoft, s'appliquent donc aux expériences IA de Microsoft Bing.
[55] https://blog.google/products/search/google-search-explicit-deep-fake-content-update/
[56] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_AI_Cleanup

- **Meta** a une politique spécifiquement intitulée « *fraude, arnaque et pratiques trompeuses* », qui englobe une série de contenus interdits dans plusieurs domaines (ex. domaine financier, identité non authentique, emploi, etc.) ;
- **Snapchat** précise que les fraudes et les pratiques trompeuses interdites incluent les contenus faisant la promotion d'escroqueries de toute nature, les programmes d'enrichissement rapide, les contenus payants ou sponsorisés non autorisés ou non présentés comme tels, le marketing à paliers multiples ou les systèmes pyramidaux et la promotion de biens ou de services frauduleux, y compris les produits ou documents contrefaits.

2. Usurpation d'identité

Classée parmi la liste des *TTPs* identifiées au sein du code de conduite contre la désinformation (cf. partie 1), l'usurpation d'identité est également interdite sur l'ensemble des grandes plateformes étudiées. Néanmoins, le niveau de précision de leurs politiques peut varier en la matière.

Par exemple, **X** précise que l'usurpation d'identité constitue une infraction à ses règles. La plateforme mentionne que, bien qu'il ne soit pas obligatoire d'afficher sa véritable identité sur son profil, le compte ne doit pas utiliser des informations qui usurpent l'identité d'autrui. X mentionne toutefois autoriser les comptes de parodie, de commentaires et de fans qui respectent ses CGU, uniquement si leur but est de discuter, de satiriser ou de partager des informations.

Pour sa part, **YouTube** fournit une explication détaillée de ce que la plateforme interdit :

- **l'usurpation d'une chaîne**, qui se produit lorsqu'une chaîne reprend le profil, l'arrière-plan ou l'apparence générale d'une autre chaîne dans le but de se faire passer pour son propriétaire. Cela ne signifie pas nécessairement que la chaîne est rigoureusement identique, mais que son intention de copier une autre chaîne est claire ;
- **l'usurpation d'identité d'une personne**, qui se matérialise par un contenu conçu pour faire croire qu'il a été publié par une autre personne.

C. Comportements inauthentiques et engagements artificiels

Les CGU des grandes plateformes sont généralement très détaillées en matière de lutte contre les comportements inauthentiques et les engagements artificiels. Cependant, ces règles restent hétérogènes d'une plateforme à l'autre, tant dans le choix de termes qui ne recouvrent pas les mêmes réalités que dans l'étendue des interdictions, reflétant ainsi des approches et des priorités différentes selon les services.

L'approche de Meta

Au titre de ses CGU, Meta interdit la création, l'utilisation et/ou la mise en réseau connecté d'éléments non authentiques (comptes, pages, groupes, etc.), notamment dans le but de tromper les utilisateurs ou de contourner les dispositifs de la plateforme (notamment les standards de la communauté et le système de signalements). En revanche, Meta demande du contexte et/ou des informations supplémentaires pour assurer le respect de ses standards de la communauté dans certains cas (ex. entités qui adoptent ou prétendent adopter un comportement non authentique coordonné ou qui se livrent ou prétendent se livrer à une ingérence étrangère).

L'approche de X

X interdit de s'engager dans une activité inauthentique qui porterait atteinte à l'intégrité de son service. À ce titre, la plateforme n'autorise pas la création, l'exploitation ou l'enregistrement massive de comptes qui ne sont pas légitimes, authentiques et transparents quant à leur source, leur identité et leur popularité. La coordination de tels comptes est également interdite, notamment pour créer une influence artificielle. Par ailleurs, l'utilisation de solutions de contournement pour dépasser les limites techniques de création de compte est prohibée[57].

L'approche de YouTube

YouTube n'autorise aucune pratique visant à augmenter artificiellement le nombre de vues, de *likes*, de commentaires ou d'autres métriques par l'utilisation de systèmes automatisés ou par la diffusion auprès d'utilisateurs non avertis. Les contenus dont l'unique but est de provoquer l'engagement des spectateurs sont également interdits. Ces règles s'appliquent aux vidéos, aux descriptions de vidéos, aux commentaires, aux diffusions en direct et à tout autre produit ou fonctionnalité de YouTube (liste non exhaustive).

L'approche de TikTok

TikTok n'autorise pas les comptes qui induisent en erreur ou qui tentent de manipuler la plateforme, ni le commerce de services qui stimulent artificiellement l'engagement ou trompent le système de recommandation. Cela inclut des comportements tels que les opérations d'influence, l'usurpation d'identité, le contenu indésirable, les faux avis et le partage de contenu piraté de manière nuisible. TikTok interdit strictement les outils d'automatisation, les scripts ou les astuces conçues pour contourner ses systèmes.
TikTok rend par ailleurs inéligible au fil « *For You* » toute action visant à inciter les utilisateurs à accroitre leur engagement.

L'approche de Microsoft (applicable à Bing et LinkedIn)

Au sein du *Microsoft Services Agreement*, qui régit notamment l'utilisation de Bing et de LinkedIn, il est stipulé que les activités frauduleuses, fausses ou trompeuses (ex. créer de faux comptes, automatiser des activités inauthentiques, générer ou partager du contenu intentionnellement trompeur, manipuler les systèmes de classement, etc.) sont interdites. Il en est de même pour le contournement des restrictions d'accès, d'utilisation ou de la disponibilité des services (ex. en tentant de *jailbreaker*[58] un système d'IA ou en effectuant un *scraping* non autorisé).

LinkedIn apporte des précisions dans les conditions d'utilisation qui lui sont propres. La plateforme n'autorise pas les logiciels tiers, y compris les robots d'indexation, les bots informatiques, les modules et extensions de navigateur qui effectuent du *web scraping*[59], modifient ou automatisent les activités sur le site de LinkedIn. Les faux comptes ne sont pas non plus autorisés, tout comme les outils ou services qui essaient de manipuler les algorithmes de LinkedIn.

[57] À noter qu'X permet aux utilisateurs de créer et/ou d'exploiter jusqu'à 10 comptes à des fins différentes et non duplicatives.
[58] Contourner des restrictions imposées par un fabricant sur un appareil.
[59] Collecte automatique de données sur des sites web au moyen de programmes ou de scripts informatiques.

Enfin, le fait de perturber le fonctionnement du service ou d'imposer une charge disproportionnée sur la plateforme (ex. spam, attaque par déni de service, virus, etc.) est également interdit.

L'approche de Snapchat

Snapchat interdit l'utilisation du service d'une manière qui pourrait déranger, perturber ou affecter négativement les utilisateurs, ou qui aurait comme conséquence d'endommager, de surcharger ou d'altérer le fonctionnement de la plateforme. Par exemple, Snapchat interdit d'utiliser des robots ou tout autre moyen informatisé, ainsi que le recours à des applications tierces, pour accéder à la plateforme.

Le fait de télécharger des virus ou d'autres codes malveillants, de compromettre, contourner ou éviter la sécurité de Snapchat est également prohibé. Dans le même ordre, le fait de tester la vulnérabilité de la plateforme est interdite.

L'approche de Google Search

Google Search n'a pas de conditions d'utilisation spécifiques similaires à celles des très grandes plateformes en ligne, qui listent des interdictions générales de comportements inauthentiques ou d'engagements artificiels. Néanmoins, quelques informations peuvent être trouvées dans les politiques dédiées au spam de Google Search (cf. *infra*, sur les contenus indésirables).

L'approche de Wikipédia

Wikipédia n'a pas de conditions d'utilisation spécifiquement dédiées aux comportements inauthentiques et aux engagements artificiels.

Cependant, les conditions d'utilisation de la *Wikimedia Foundation* et les politiques de la communauté interdisent certains comportements problématiques comme :

- le vandalisme, défini comme la modification d'une page de manière intentionnellement perturbatrice ou malveillante ;
- le fait de tester les vulnérabilités des systèmes ou des réseaux techniques de Wikipédia ;
- le fait d'accéder, sans autorisation, aux systèmes informatiques de Wikipédia, de les utiliser et d'en altérer le fonctionnement.

D. Contenus indésirables (spams)

Toutes les grandes plateformes étudiées ont des politiques permettant de limiter les risques de spams sur leurs services.

Certains d'entre eux distinguent deux catégories[60] :

[60] Ces définitions figurent dans les CGU d'X.

- **les spams d'engagement** : utilisation non authentique des fonctionnalités d'engagement de la plateforme pour générer artificiellement du trafic ou perturber l'expérience des utilisateurs ;
- **les spams de contenu** : contenu publié de manière massive, dupliquée, non pertinente ou non sollicitée qui perturbe l'expérience des utilisateurs.

De manière générale, ils interdisent systématiquement les techniques de manipulation destinées à perturber le bon fonctionnement de leur service, attirer l'attention ou à inciter à interagir avec du contenu (Meta, X, YouTube, TikTok), comme les publications de type « *piège à clics* » (TikTok), le *copypasta* (X), les titres, les miniatures, commentaires ou descriptions trompeuses (YouTube, Microsoft Bing et LinkedIn), ainsi que la modification de liens pour rediriger les utilisateurs vers des sites externes ou des services tiers (X, YouTube, Snapchat, Google Search).

Les grandes plateformes prohibent également la création, le partage ou la vente de comptes, pages, groupes, événements, privilèges ou interactions de manière excessive ou artificielle (Meta, X). Les comportements visant à manipuler les indicateurs de popularité ou le classement algorithmique (Microsoft Bing, X) – par exemple par des abonnements/désabonnements en masse ou « *follow churn*[61] », coordination d'échanges d'engagement ou utilisation de services tiers pour gonfler artificiellement l'audience – sont également interdits (X). Il en est de même des pratiques qui génèrent du trafic ou redirigent les utilisateurs de manière non transparente.

Toutes les grandes plateformes étudiées mettent l'accent sur la sécurité et l'authenticité : l'hameçonnage (Microsoft Bing et LinkedIn, X), la diffusion de logiciels malveillants (Google Search), l'utilisation abusive de l'automatisation (X) et la création massive de contenus sans valeur ajoutée (Meta, X, YouTube, TikTok, Snapchat, Google Search, Wikipédia) sont explicitement interdits.

II. L'apport des systèmes de détection automatisés

Face à l'évolution constante des *TTPs* employées par des acteurs malveillants, les grandes plateformes ont renforcé leurs capacités de détection automatisée des comportements inauthentiques. Ces systèmes jouent un rôle central dans la lutte contre la manipulation de l'information, en complément des équipes de modération humaine (cf. *infra*). Cet effort s'inscrit dans une dynamique permanente d'adaptation, les acteurs malveillants testant en continu de nouvelles stratégies pour contourner les mécanismes de détection et l'application des conditions d'utilisation des plateformes (ex. recours, à l'utilisation détournée d'émojis, de hashtags, de menaces implicites ou de langage codé).

Meta illustre particulièrement cette dynamique d'adaptation : l'entreprise dispose de mécanismes proactifs visant à identifier de nouveaux schémas de comportements malveillants et à les intégrer dans son système de détection automatisé. Lorsque certains types de contenus ou de comportements apparaissent plus fréquemment, ses modèles sont entraînés et améliorés en conséquence. Les systèmes de détection automatisés de Meta identifient également des réseaux coordonnés cherchant à réintégrer ses services après avoir été supprimés. Les travaux de Meta en la matière

[61] L'abonnement/désabonnement en masse ou « *follow churn* » est une technique visant à suivre pour à se désabonner immédiatement d'un grand nombre de comptes dans le but d'augmenter son propre nombre d'abonnés.

sont documentés dans ses rapports réguliers sur les menaces adverses[62]. Pour la détection des *deepfakes* en particulier, les processus de Meta évoluent chaque année.

Pour sa part, **Google Search** s'appuie sur des technologies d'IA avancées pour renforcer sa capacité à identifier et supprimer les pages frauduleuses de ses résultats de recherche. Ses algorithmes analysent de nombreux signaux afin de garantir la légitimité des contenus proposés aux utilisateurs et à les protéger contre des sites nuisibles. Par ailleurs, l'outil anti-spam de Google Search, nommé « *Spam Brain* »[63], permet au moteur de recherche de prendre des mesures proactives contre le spam à très grande échelle. En matière d'intégrité civique, le moteur de recherche utilise des classificateurs spécifiques pour identifier les requêtes liées aux élections, signalant aux systèmes l'importance de fournir des informations fiables, récentes et issues de sources faisant autorité.

L'apprentissage automatique est également au cœur des systèmes de **YouTube**, que la plateforme considère comme particulièrement efficace pour identifier les tendances et repérer des contenus similaires à ceux ayant déjà fait l'objet d'une suppression, avant même qu'ils soient consultés par les utilisateurs. YouTube continue d'investir dans l'amélioration de ces systèmes et s'appuie sur des évaluateurs humains et sur l'apprentissage automatique pour les entraîner à l'utilisation de nouvelles données. Les équipes d'ingénierie de YouTube mettent régulièrement à jour ces systèmes et cherchent à exploiter une combinaison encore plus ciblée de classificateurs, de mots-clés dans un nombre plus important de langues et d'informations provenant d'analystes régionaux afin d'identifier des récits que son classificateur principal ne détecte pas. En outre, afin de mesurer les progrès dans la rapidité des retraits de vidéos préjudiciables ou illégales, YouTube a également développé un nouveau système métrique appelé *Violative View Rate* (VVR) qui permet d'évaluer la visibilité des contenus avant qu'ils soient retirés et d'analyser les préjudices potentiels causés par la diffusion de tels contenus avant leur retrait.

Microsoft a également mis en place un ensemble de mesures d'atténuation des risques reposant sur des classificateurs, mais aussi des filtres de contenu et des techniques de *méta-prompting*[64] afin de réduire les risques de préjudices et d'abus liés à ses services, notamment ceux intégrant de l'IA générative. Les filtres de contenus et les classificateurs permettent notamment d'empêcher la mise en avant de contenus considérés comme de faible autorité et permet de lutter contre les *TTPs*. Les algorithmes et les systèmes de classement de Microsoft intègrent des renseignements sur les menaces liées aux élections et à la manipulation de l'information, ainsi que des données électorales localisées pour orienter les utilisateurs vers des sources officielles.

Comme toutes les autres grandes plateformes précitées, **LinkedIn** investit continuellement dans de nouvelles technologies pour lutter contre les comportements inauthentiques sur la plateforme. Elle utilise notamment des algorithmes de réseau avancés capables d'identifier des communautés de faux comptes[65] sur la base de similitude de contenus et de comportements, des algorithmes de vision par ordinateur[66]

[62] https://transparency.meta.com/metasecurity/threat-reporting/
[63] https://spambrain.com/
[64] La technique de « *meta-prompting* » consiste à donner des méta-instructions, qui s'imposent par défaut à toutes les requêtes des usagers et encadrent les réponses du modèle d'IA, pour guider son comportement, notamment pour que le système se comporte conformément aux principes d'IA (tels que définis par Microsoft dans ce cadre) et aux attentes des utilisateurs.
[65] Les défenses automatisées de LinkedIn ont bloqué 97,1 % des faux comptes détectés entre juillet et décembre 2024, les 2,9 % restants ayant été bloqués grâce à des enquêtes manuelles. 99,7 % des faux comptes ont été bloqués de manière proactive, avant même qu'un membre ne les signale.
[66] La vision par ordinateur est un sous-domaine de l'intelligence artificielle qui permet aux ordinateurs et aux systèmes de reconnaitre automatiquement les images et les vidéos et de les décrire avec précision.

et de traitement du langage naturel pour détecter des éléments générés par IA, des anomalies de comportements à risque et des séquences d'activités associées à l'automatisation abusive.

De son côté, **TikTok** indique avoir mis en place des techniques dites de dispersion[67] fondées sur l'apprentissage automatique, afin d'éviter la recommandation répétée de vidéos similaires portant sur des thèmes sensibles. Ainsi, même lorsque ces contenus ne violent pas directement les règles de la plateforme, cette approche vise à réduire les risques liés à une approche prolongée ou cumulative.

Enfin, **Snapchat** dispose aussi de procédures de modération proactives, en particulier basées sur la détection de mots clés.

***

Si les grandes plateformes présentent des CGU en matière de lutte contre les techniques de manipulation (cf. détail des CGU pertinentes en annexe 1), et des dispositifs de détection présentés comme étant efficaces, de nombreuses études (de VIGINUM, des organisations de la société civile et des chercheurs) documentent régulièrement des opérations d'ingérence étrangère ou de manipulation de l'information observables sur leurs services, qui laissent à s'interroger sur l'efficacité réelle des CGU et leur mise en œuvre effective. Ces travaux questionnent également sur la fiabilité et la performance des systèmes de détection automatisés, ainsi que sur le temps de réponse aux signalements que les grandes plateformes reçoivent. À ce titre, des chercheurs de *Providus* (groupe de réflexion letton) ont mené un projet de surveillance[68] visant plusieurs plateformes (TikTok, YouTube, Facebook et Telegram), qui leur a permis de démontrer des manquements en matière de réponse aux signalements. Si certaines plateformes les ont ignorés, d'autres ont répondu par des messages automatiques : « *Aucune infraction constatée* ». Pour autant, un mécanisme de notification et d'action est prévu par l'article 16 du RSN afin de permettre aux utilisateurs de signaler les contenus suspectés d'être illicites. À noter que l'article 16.5 précise également que « *le fournisseur notifie également, dans les meilleurs délais, […] sa décision concernant les informations auxquelles la notification se rapporte, tout en fournissant des informations sur les possibilités de recours à l'égard de cette décision* ».

III. L'importance des approches humaines et collaboratives

Si les outils automatisés et les systèmes de détection constituent un pilier essentiel dans la lutte contre les *TTPs* employées par les acteurs malveillants, ils ne peuvent à eux seuls répondre à la complexité et à l'évolution constante des risques pouvant avoir un impact sur l'espace informationnel en ligne. Les grandes plateformes reconnaissent ainsi que la protection de l'intégrité de leurs services repose également sur des approches humaines et collaboratives, capables d'apporter du contexte, de l'expertise sectorielle et une compréhension fine des dynamiques sociales et informationnelles.

À cet égard, leurs équipes de modération humaine jouent un rôle central, en ce qu'elles permettent d'interpréter des contenus ambigus, de détecter des récits propres à des contextes locaux spécifiques ou encore d'identifier des schémas de manipulation sophistiqués qui échappent parfois aux systèmes automatisés. En parallèle, les partenariats avec des acteurs externes – vérificateurs de faits, chercheurs, organisations de la société civile, médias et autres entreprises technologiques –

[67] Une technique dite de dispersion désigne un mécanisme algorithmique visant à limiter la concentration et la répétition excessive de contenus similaires dans les flux de recommandation à un utilisateur.
[68] "No Violations Found." Europe's Digital Safety Law Fails When Users Report Content – EDMO

contribuent au renforcement de la capacité des plateformes à anticiper les menaces. À noter que certaines enceintes multilatérales, comme la *taskforce* du code de conduite contre la désinformation, offrent également un cadre de dialogue et de coopération réunissant certaines grandes plateformes, acteurs industriels et organisations de la société civile.

A. Partenariats avec des vérificateurs de faits

La collaboration avec les vérificateurs de faits constitue un levier essentiel pour compléter les capacités internes des plateformes, en apportant une expertise spécialisée, une méthodologie rigoureuse et un regard externe sur l'exactitude des contenus circulant sur les services des grandes plateformes. Ceux-ci jouent en effet un rôle central dans l'identification, l'analyse et la contextualisation des contenus faux ou trompeurs, en particulier lorsqu'ils portent sur des enjeux sensibles tels que les processus électoraux, la santé publique ou les crises géopolitiques. Leur intervention permet non seulement d'évaluer la véracité des affirmations diffusées, mais aussi d'alimenter les dispositifs de modération des grandes plateformes. De ce fait, certains d'entre eux ont développé des systèmes organisationnels structurés pour intégrer ces expertises externes dans leurs processus décisionnels.

À ce titre, **TikTok** s'appuie notamment sur un réseau de partenaires de vérification des faits accrédités par l'*International Fact-Checking Network* (IFCN) afin de limiter la diffusion de contenus faux ou trompeurs. En France, la plateforme travaille en étroite collaboration avec l'Agence France-Presse (AFP) depuis 2020, laquelle peut lui signaler des contenus ou évaluer ceux que TikTok lui transmet. TikTok dispose par ailleurs de modérateurs ayant accès un référentiel mondial d'affirmations précédemment vérifiées par des partenaires indépendants (ce qui est aussi le cas de **LinkedIn**[69]), et a également mis en place des équipes « *de terrain* », qui collaborent avec des experts locaux afin d'intégrer les contextes culturels, linguistiques et politiques propres à chaque territoire.

Meta a également institutionnalisé des partenariats à l'échelle de l'Union Européenne avec des vérificateurs de faits certifiés par l'*European Fact-Checking Standards Network* (EFCSN). Dans ce cadre, Meta a notamment mené des ateliers avec l'EFCSN, pour améliorer les compétences et les capacités de la communauté européenne des vérificateurs de faits à démystifier et lutter contre la désinformation générée par l'IA et à faciliter la mise en place de normes communes pour la vérification de tels contenus. Néanmoins, l'arrêt des programmes de vérification des faits aux États-Unis, annoncée en janvier 2025 par Mark Zuckerberg, suscite une vive incertitude quant au renouvellement du programme européen pour les années à venir[70]. Selon une étude du *Center for Countering Digital Hate* (CCDH)[71] publiée le 24 février 2025, l'abandon par Meta de son programme de soutien aux vérificateurs de faits pourrait se traduire par un « *raz-de-marée* » de contenus préjudiciables aux utilisateurs sur ses services.

En s'appuyant sur ces réseaux de vérificateurs de faits, les grandes plateformes renforcent la crédibilité et la cohérence de leurs politiques, d'autant plus que ces partenariats participent à une meilleure compréhension des phénomènes de manipulation de l'information sur les services numériques.

[69] À noter que LinkedIn collabore également avec des organisations de presse lorsque la véracité des contenus qui lui sont signalés ne peut être confirmée.
[70] *Cf.* Partie 4 – Recommandations.
[71] More Transparency and Less Spin — Center for Countering Digital Hate | CCDH

B. Partenariats avec des évaluateurs externes : l'exemple des « *Google Search Quality Raters* »

Les *Google Search Quality Raters* (évaluateurs externes de la qualité de la recherche) sont des évaluateurs humains mandatés par Google pour apprécier la qualité des résultats affichés par son moteur de recherche et évaluer l'efficacité de la recommandation. S'appuyant sur des lignes directrices fournies par Google[72], ils sont chargés de mesurer les performances de ses différents systèmes de classement des résultats de recherche, bien que leurs évaluations n'influencent pas directement le classement.

À noter que la publication des lignes directrices par Google contribue à une forme de transparence méthodologique, même si le détail des pondérations algorithmiques reste confidentiel.

C. Collaboration avec des experts de la société civile et du monde académique

Certaines grandes plateformes soulignent l'importance d'une approche fondée sur la collaboration avec des experts spécialisés. **Snapchat**, par exemple, travaille avec des experts et des *think thanks* de renom comme le *Digital Forensic Research Lab* (DFRLab[73]) de l'*Atlantic Council* et a également participé à des initiatives de recherche collaborative comme le *Election Integrity Partnership*[74], un partenariat avec des organisations de la société civile et des centres de recherche dédié à la surveillance des atteintes aux processus démocratiques en ligne. **LinkedIn** entretient également des liens étroits avec des organisations de la société civile pour mieux comprendre les moyens visant à abuser de sa plateforme et les risques pour ses membres, tout comme **TikTok** qui, dans le cadre de son programme transversal sur l'intégrité des élections, cherche à améliorer ses politiques et processus afin de comprendre les contextes locaux et d'obtenir des informations faisant autorité.

De son côté, **Google** a participé en 2024 au *Trust & Safety Forum* à Lille, réunissant plus de 30 participants dont des universitaires, pour explorer les cadres de sécurité dans la conception des produits et les contraintes de mise en œuvre, notamment eu égard à la manipulation de l'information. La même année, Google a lancé le premier cycle du *Google Academic Research Award* (GARA), un programme de soutien « *à la recherche innovante en informatique et en technologies susceptibles de générer des applications concrètes* »[75]. Ce projet s'est poursuivi en 2025, Google ayant annoncé en octobre soutenir 56 projets menés par 84 chercheurs de 12 pays différents. Google collabore également avec le projet Lumen, un projet de recherche indépendant sur la disponibilité des contenus en ligne.

Microsoft collabore aussi avec des organisations spécialisées telles que l'*Institute for Strategic Dialogue* (ISD) et NewsGuard. Par exemple, dans le cadre du partenariat avec ce dernier, Microsoft met à disposition des utilisateurs de son navigateur Edge l'accès gratuit aux évaluations et aux « *étiquettes nutritionnelles* » de Newsguard, et intègre ces données dans certains services comme Bing, afin d'aider les utilisateurs à mieux repérer les sources fiables et à lutter contre la manipulation de l'information en ligne.

[72] Lignes directrices de Google pour les évaluateurs de la qualité de la recherche
[73] https://www.atlanticcouncil.org/programs/digital-forensic-research-lab/
[74] https://www.eipartnership.net/2020
[75] https://blog.google/outreach-initiatives/google-org/congratulations-to-the-2025-google-academic-research-award-recipients/

D. Partenariats industriels et technologiques

La lutte contre les techniques de manipulation de l'information repose également sur des partenariats industriels et technologiques permettant aux grandes plateformes de mutualiser les expertises, les outils et les bonnes pratiques face à des menaces complexes et évolutives. Ces coopérations favorisent une approche coordonnée pour anticiper les risques liés aux nouveaux usages technologiques, en particulier ceux associés à l'intelligence artificielle et aux contenus synthétiques, tout en renforçant la capacité des plateformes à prévenir les manipulations de l'information à grande échelle.

Dans ce cadre, plusieurs grandes plateformes jouent un rôle actif dans des initiatives collectives visant à promouvoir un développement responsable des technologies. **Google** est ainsi fondateur ou participant à différentes initiatives industrielles récentes, parmi lesquelles le *Frontier Model Forum*, organisme axé sur la promotion d'un développement sûr et responsable des modèles d'IA de pointe, ou encore le partenariat sur l'IA.

Focus sur le partenariat sur l'IA

Le partenariat sur l'IA[76] est un forum de coopération qui réunit des entreprises technologiques, des organisations de la société civile, des chercheurs, des médias et des institutions académiques. Sa mission est de promouvoir un développement et un usage responsable et éthique de l'IA.

Les grandes plateformes suivantes en sont membres : Google, TikTok, Microsoft et Meta.

Ce partenariat permet notamment :
- d'anticiper et d'analyser les risques liés à l'IA, notamment en matière de manipulation de l'information et de médias synthétiques ;
- d'élaborer des principes, cadres et bonnes pratiques volontaires partagés par les acteurs du secteur ;
- de favoriser le dialogue entre industrie, recherche et société civile.

Parmi ses travaux majeurs figure le document intitulé « *pratiques responsables pour les médias synthétiques : un cadre d'action collective* », qui propose un ensemble de recommandations visant à soutenir le développement et le déploiement responsable des médias synthétiques.

Par ailleurs, certaines grandes plateformes (**TikTok**, **Meta**, **Google**[77] ou encore **Microsoft**) soutiennent ou participent activement à la norme industrielle C2PA (*Coalition for Content Provenance and Authenticity*), qui vise à certifier la provenance et l'authenticité des contenus en ligne au moyen de métadonnées sécurisées retraçant l'origine d'un contenu, les conditions de sa création et les éventuelles modifications qu'il a subies.

Les partenariats peuvent également être bilatéraux et opérationnels : à ce titre, **LinkedIn** collabore avec des entreprises pour recevoir des indicateurs liés à la création

[76] https://syntheticmedia.partnershiponai.org/#landing
[77] Par exemple, en septembre 2024, Google a annoncé l'intégration de la dernière version de la norme technique C2PA à la fonctionnalité « *À propos de cette image* » de Google Search.

de faux comptes par des acteurs soutenus par des États et échanger des informations sur les *TTPs* et les acteurs de menaces persistantes avancées[78].

Présentation de la *taskforce* du code de conduite contre la désinformation

L'engagement 37 du code institue une *taskforce* à laquelle participent tous les signataires, tels que des grandes plateformes (Facebook, Instagram, Google Search, YouTube, Microsoft Bing, LinkedIn, TikTok), des acteurs industriels (ex. Adobe, Double Verify) et des organisations de la société civile (ex. AI Forensics, Reporters sans frontières, Democracy Reporting International). Cette *taskforce,* présidée par la Commission européenne, inclut également des représentants de l'EDMO, du *Media Board*[79] (l'Arcom y participe à ce titre) et du Service européen pour l'action extérieure (SEAE).

Les signataires se réunissent périodiquement pour échanger sur certains sujets en lien avec le code, qu'ils soient thématiques ou organisationnels (ex. souscription aux engagements, définition d'une liste commune de *TTPs*, mise à jour du code, etc.).

Ce type de dispositif, soutenu par l'Arcom, permet d'instaurer un cadre de réponse collective et coordonnée et offre ainsi un environnement précieux pour renforcer la compréhension commune des enjeux liés à la manipulation de l'information en ligne.

Néanmoins, l'Arcom note un désengagement progressif des grandes plateformes qui en sont signataires. Ce constat inquiétant est également dressé par le réseau européen des normes de vérification des faits (EFCSN), dans une étude publiée en septembre 2025[80] intitulée « *moment de vérité pour le code de conduite contre la désinformation* ».

IV. Les actions éducatives et de sensibilisation en matière de citoyenneté du numérique

Les grandes plateformes peuvent également agir via les actions éducatives et de sensibilisation en matière de citoyenneté numérique qu'elles conceptualisent et rendent disponibles.

En fournissant aux utilisateurs des clés de compréhension de l'espace informationnel en ligne et des éléments visant à se prémunir des comportements malveillants et des opérations de manipulation de l'information, ces actions visent à renforcer leur esprit critique et contribuent à accroître la résilience des utilisateurs. Les initiatives générales en matière de citoyenneté et d'éducation au numérique sont ainsi complémentaires aux modules destinés à comprendre le fonctionnement des IA, et peuvent s'appliquer à des cas concrets, comme le climat ou les élections[81].

[78] Une menace persistante avancée (ou « *Advanced Persistent Threat* ») désigne un acteur ou un groupe de cybermenace hautement sophistiqué, généralement soutenu ou lié à un État, qui mène des cyberattaques ciblées et persistantes au sein d'un réseau.
[79] Comité européen pour les services de médias, institué par le règlement 2024/1083 établissant un cadre commun pour les services de médias dans le marché intérieur (EMFA).
[80] The Moment of Truth for the Code of Conduct on Disinformation – European Fact-Checking Standards Network (EFCSN)
[81] https://www.arcom.fr/se-documenter/etudes-et-donnees/etudes-bilans-et-rapports-de-larcom/rapport-sur-leducation-aux-medias-linformation-et-la-citoyennete-numerique-exercice-2024-2025 : les initiatives des plateformes sont d'autant plus crédibles qu'elles fournissent des données sur leur impact, tel que le nombre d'utilisateurs touchés.

A. Exemples d'initiatives en matière d'éducation au numérique

Les initiatives en matière de citoyenneté et d'éducation au numérique visent à donner aux utilisateurs les outils et les connaissances nécessaires pour naviguer de manière critique et responsable dans l'environnement numérique. Elles permettent notamment de développer des compétences de discernement, de comprendre le fonctionnement des plateformes, et d'acquérir une certaine vigilance vis-à-vis des contenus diffusés.

Dans ce cadre, **Google Search** a lancé plusieurs programmes éducatifs, dont le programme *Be Internet Awesome*[82], destiné aux enfants, qui enseigne les bases de la citoyenneté numérique et de la sécurité, afin qu'ils puissent explorer l'environnement en ligne sereinement et identifier les contenus faux et trompeurs. En complément, la campagne *Hit Pause*[83] lancée sur **YouTube** en novembre 2022 vise à renforcer les compétences des utilisateurs en matière d'éducation aux médias. Ce projet, porté par l'équipe *Trust and Safety* de YouTube, est diffusé sur une chaîne dédiée et encourage les utilisateurs à prendre du recul face à un contenu et à réfléchir avant d'agir à son sujet, notamment pour le repartager.

Meta déploie également des initiatives d'éducation au numérique dans l'Union européenne, en proposant un centre d'éducation accessible aux utilisateurs qui fournit des ressources sur la désinformation et accompagne les adolescents dans leur navigation en ligne.

B. Outils pour comprendre l'intelligence artificielle

Dans un environnement numérique ou l'IA devient omniprésente, les grandes plateformes ont mis en place des mesures d'étiquetage permettant aux utilisateurs de mentionner que leurs contenus ont été générés ou modifié par un système d'IA[84]. Ces étiquettes informatives contribuent à renforcer la transparence et la confiance dans l'information consultée par les utilisateurs. En cas de signalement d'un contenu généré par IA et non étiqueté comme tel, la plateforme effectue d'elle-même cet étiquetage.

Au-delà de cette fonctionnalité, certaines grandes plateformes proposent également des outils éducatifs et des ressources pédagogiques expliquant le fonctionnement des modèles d'IA, leurs limites et les biais ou erreurs possibles. Leur objectif est notamment de développer la compréhension critique des utilisateurs afin qu'ils puissent identifier des contenues synthétiques ou manipulés.

Meta a déployé plusieurs actions significatives dans ce domaine, en particulier en 2024 :

- l'entreprise a notamment collaboré avec l'EFCSN pour lancer une campagne d'éducation aux médias visant à permettre aux utilisateurs d'identifier des contenus générés par l'IA[85] ;
- Meta France a produit avec l'AFP un « *Réel* »[86] mettant en scène Thomas Pesquet, destiné à sensibiliser le public aux images et vidéos trompeuses diffusées en ligne ;
- Meta s'est associé à Génération Numérique pour diffuser des vidéos pédagogiques vulgarisant l'IA générative ;

[82] https://beinternetawesome.withgoogle.com/fr_all/interland
[83] https://www.youtube.com/hitpause
[84] Il s'agit d'une obligation au titre de l'article 50 du règlement 2024/1689 du 13 juin 2024 sur l'intelligence artificielle.
[85] https://efcsn.com/news/2024-04-18_efcsns-new-project-for-identifying-ai-generated-and-digitally-altered-content/
[86] Un « *Réel* » désigne un format de courte vidéo verticale.

- l'entreprise a aussi développé des « *cartes de système d'IA* »[87] fournissant aux utilisateurs des explications sur le fonctionnement de l'IA dans ses produits et sur l'impact de leur comportement numérique sur les contenus qui leur sont proposés.

TikTok a également élaboré des ressources éducatives avec Mediawise et WITNESS, destinées à aider les utilisateurs à repérer les contenus générés par l'IA. Ces campagnes de sensibilisation ont permis d'atteindre 80 millions d'utilisateurs dans le monde, dont plus de 9 millions en France.

C'est également le cas de Microsoft, qui a lancé une initiative d'éducation aux médias centrée sur la prévention de l'usage trompeur de l'IA, en particulier en contexte électoral, ainsi que la campagne « *Be Informed, Not Misled* » dans le cadre du *News Literacy Project*. **Bing** propose par ailleurs des foires aux questions, des pages d'aide et d'autres sources d'information accessibles au public, pour informer les utilisateurs sur la nature des expériences de recherche pilotées par l'IA.

Google a choisi d'appuyer financièrement des organisations capables de développer des ressources et des projets en lien avec l'IA. Au second semestre 2024, il a financé 10 millions de dollars à la Fondation Raspberry Pi pour élargir l'accès à « *Experience AI* », un programme éducatif cocréé avec Google DeepMind. Dans ce cadre, des modules[88] axés sur les fondements et la sécurité de l'IA ont été développés.

C. L'éducation au climat par TikTok

TikTok se distingue des autres grandes plateformes en étant la seule à disposer de conditions d'utilisation spécifiquement dédiées à la lutte contre la désinformation climatique. Cette reconnaissance explicite du risque informationnel lié aux enjeux climatiques s'accompagne d'une stratégie proactive visant à renforcer l'éducation climatique de ses utilisateurs, à promouvoir des sources fiables et à encourager une participation éclairée aux débats environnementaux.

Dans ce cadre, TikTok a lancé une initiative d'un million de dollars, en partenariat avec *Verified for Climate*, destinée à lutter contre la désinformation climatique[89]. Cette initiative a réuni des scientifiques et des experts de confiance et visait à sélectionner des créateurs de TikTok afin qu'ils produisent des contenus éducatifs fondés sur des données scientifiques, tout en encourageant l'action climatique au sein de la communauté de la plateforme.

À l'occasion de la COP28, TikTok a également lancé sa campagne annuelle #ClimateAction, visant à encourager les communautés d'utilisateurs à s'informer, à échanger et à s'engager sur les enjeux climatiques. Cette campagne, qui s'inscrit dans un engagement de long terme[90], s'est poursuivi lors de la COP29[91], à laquelle TikTok avait participé activement en tant que partenaire stratégique.

[87] https://transparency.meta.com/features/explaining-ranking
[88] https://experience-ai.org/en/units/?utm_source=blog&utm_medium=organic&utm_campaign=expai-Q1&utm_id=Experience+AI&utm_content=impactblog1
[89] https://newsroom.tiktok.com/advancing-our-commitment-to-sustainability-and-climate-literacy-at-cop-28?lang=en
[90] La campagne a été lancée lors de la COP26.
[91] https://www.tiktok.com/for-good/cop29/

Enfin, en 2024, TikTok s'est associé au Centre Mary Robinson pour lancer la « *Youth Climate Leaders Alliance* »[92], un programme destiné aux 18-30 ans et visant à leur apporter des compétences et des outils dans la lutte contre le changement climatique.

Néanmoins, la Fondation Maldita.es et AI Forensics ont révélé, en novembre 2025, que TikTok n'avait pas réussi à modérer la désinformation concernant les inondations à Valence en 2024, et avait même amplifié ce type de contenu[93]. L'Arcom s'interroge ainsi sur la mise en œuvre effective de cette politique par TikTok.

D. Initiatives autour des élections

Les initiatives et sensibilisations en matière électorale mises en œuvre par les grandes plateformes constituent une mesure essentielle d'atténuation des risques systémiques relatifs au discours civique, aux processus électoraux et à la sécurité publique, tels que visés à l'article 34 du RSN. En renforçant la capacité des utilisateurs à comprendre les enjeux des processus électoraux et à identifier les comportements inauthentiques qui pourraient les impacter, ces actions contribuent directement à la résilience de la société et contribuent à renforcer l'intégrité des scrutins.

Dans ce cadre, **Meta** a notamment collaboré avec le Forum européen des personnes handicapées (FEPH) pour lancer une initiative d'éducation aux médias relative à l'accessibilité des élections, visant à garantir que les informations électorales soient compréhensibles et accessibles par tous les citoyens.

Pour sa part, **TikTok** a développé plusieurs actions structurantes :

- en partenariat avec l'AFP, TikTok a développé une campagne d'éducation aux médias[94] afin de sensibiliser les utilisateurs de la plateforme, et notamment les plus jeunes, aux enjeux de la désinformation ;
- TikTok a également mené une dizaine de campagnes temporaires d'éducation aux médias et à l'information sur l'intégrité des élections avant les élections européennes de 2024, la plupart en collaboration avec ses partenaires de vérification des faits ;
- la plateforme a également organisé des sessions de conférence sur les élections ;
- TikTok a aussi continué à développer son centre de transparence en y intégrant notamment un centre d'intégrité des élections et des rapports dédiés aux opérations d'influence secrètes.

Snapchat, de son côté, a privilégié des formats interactifs et pédagogiques. En coopération avec le ministre de l'Intérieur néerlandais, la plateforme a développé un outil permettant aux utilisateurs d'améliorer leurs connaissances électorales au moyen de quizz vrai/faux. La plateforme fait également des campagnes de sensibilisation en amont des élections.

Microsoft a adopté une approche ciblée sur les risques liés à l'IA trompeuse en contexte électoral et déploie notamment des campagnes de sensibilisation et de formation à destination des partis politiques et des équipes de campagne dans l'Union européenne, afin de les aider à identifier et signaler les usages trompeurs de l'IA et les manipulations de l'information. À cette fin, l'entreprise a une page web dédiée au signalement de *deepfakes* électoraux, accessible et localisée dans l'ensemble de l'Union européenne.

[92] https://newsroom.tiktok.com/tiktok-partners-with-the-mary-robinson-centre-for-new-youth-climate-leaders-alliance?lang=en-IE
[93] YouTube est aussi concernée par cette étude.
[94] https://newsroom.tiktok.com/fr-fr/tiktok-afp-luttent-contre-la-desinformation

Enfin, **Google** soutient également des initiatives de sensibilisation électorale, notamment à travers son appui à *Election24Check*[95], une coalition de plus de 40 médias et organisations de vérification des faits, qui visait à renforcer la collaboration transnationale dans la détection et le rétablissement des faits relevant de désinformation et visant les élections européennes 2024.

***

Il peut être observé que les grandes plateformes mettent en œuvre une diversité de mesures d'atténuation pour protéger l'intégrité de leurs services et lutter contre les *TTPs* et la désinformation. Toutefois, ces acteurs reconnaissent un risque résiduel sur leurs services, qui persiste malgré l'ensemble des mesures mises en œuvre. Les dispositifs qu'ils déploient restent notamment confrontés à l'évolution constante des tactiques adverses et à certains risques d'erreur. Si ces phénomènes sont – partiellement – documentés dans les rapports publiés en application du RSN, l'Arcom rappelle que la transparence des grandes plateformes constitue un levier indispensable dans la lutte contre les *TTPs*. Elle permet notamment aux chercheurs, aux autorités publiques et à la société civile de mener des analyses plus approfondies sur les vulnérabilités des plateformes et d'anticiper en conséquence les évolutions de ces techniques. Sans ouverture suffisante sur les outils utilisés, les processus décisionnels et les résultats obtenus, les dispositifs de lutte risquent de demeurer partiels, contestés ou inefficaces, limitant ainsi leur capacité à répondre durablement aux dynamiques complexes de la manipulation de l'information en ligne.

[95] https://blog.google/around-the-globe/google-europe/fighting-misinformation-online-elections/

PARTIE 4. RECOMMANDATIONS VISANT À RENFORCER LA LUTTE CONTRE LES TECHNIQUES DE MANIPULATION

À la lumière du bilan réalisé sur les moyens et mesures mis en place par les grandes plateformes pour protéger l'intégrité de leurs services et lutter contre les *TTPs*, l'Arcom propose une série de recommandations à destination à la fois de la Commission européenne, des plateformes, des pouvoirs publics nationaux et des organisations de la société civile. Ces recommandations visent à renforcer la prévention, la détection et l'atténuation de ces pratiques et comportements interdits pour lutter contre les risques de manipulation de l'information en ligne dans le cadre du RSN.

Alors que de fortes attentes s'expriment dans les États membres en faveur d'une application pleinement effective du RSN, les coordinateurs pour les services numériques (CSN) ont notamment pour mission de faciliter la mobilisation et la participation des acteurs nationaux – autorités publiques, monde académique et société civile – dans la supervision des obligations d'identification et d'atténuation des risques systémiques induits par les services des grandes plateformes.

Dans ce contexte, l'Arcom appelle en premier lieu la Commission européenne à poursuivre avec diligence les enquêtes en cours sur de potentiels manquements aux obligations de lutte contre les risques systémiques, et notamment en termes de manipulation de l'information en ligne.

Les décisions rendues par la Commission européenne en décembre 2025[96] en ce qui concerne notamment TikTok et X traduisent des premières avancées encourageantes. Il est important qu'à l'avenir, les CSN soient mieux associés aux procédures ouvertes ou susceptibles de l'être par la Commission, notamment dans le cadre du Comité européen des services numériques (*DSA Board*).

Les recommandations de l'Arcom s'inscrivent en cohérence avec la Stratégie nationale de lutte contre les manipulations de l'information d'origine étrangère (2026-2030), publiée par le SGDSN.

Recommandation 1 : Renforcer l'effectivité du RSN en matière de lutte contre la manipulation de l'information

Acteurs visés : Commission européenne, pouvoirs publics nationaux

Sous-recommandation 1.1 : Mieux tirer profit des outils et dispositifs d'accès aux données de contenus et d'audience au titre du RSN

La pleine mobilisation des outils prévus par le RSN pour accroître la transparence des plateformes en ligne est essentielle pour permettre aux coordinateurs pour les services numériques, aux autorités nationales compétentes, à la société civile et aux chercheurs d'observer, en conditions réelles d'utilisation, le fonctionnement des plateformes en ligne, d'analyser les mesures déployées pour lutter contre les *TTPs* et de conduire des

[96] V. Commission européenne, communiqués du 5 décembre 2025 concernant TikTok et X.
https://ec.europa.eu/commission/presscorner/detail/fr/ip_25_2940
https://digital-strategy.ec.europa.eu/fr/news/commission-fines-x-eu120-million-under-digital-services-act

analyses fiables et détaillées de pratiques susceptibles d'enfreindre le règlement. Il s'agit en particulier de :

- **l'accès aux données individuelles publiques et non-publiques par les chercheurs (article 40**), qui doit permettre de corriger les asymétries d'information afin de documenter plus en profondeur l'évolution et la gravité des risques systémiques et établir des mesures d'atténuation ;

- **la transparence renforcée des grandes plateformes en matière de publicité numérique (article 39)**, qui doit permettre de rendre visible et connu de tous le recours à des ciblages publicitaires à des fins politiques (au sens du droit européen), de confirmer la réalité de l'interdiction de ce type de ciblage annoncée par certaines grandes plateformes et de détecter les usages à des fins d'escroquerie qui utilisent des contenus trompeurs, participant à la désinformation, pour piéger leurs victimes.[97]

Afin de renforcer l'effectivité de ces outils, l'Arcom encourage :

- **l'adoption des lignes directrices prévues à l'article 39 du RSN pour harmoniser la structure, l'organisation et le fonctionnement des registres publicitaires** ;

- **la poursuite des travaux visant à articuler de manière cohérente le RSN avec le règlement relatif à la transparence et au ciblage de la publicité à caractère politique** du 13 mars 2024[98], notamment en précisant :
  - la portée des obligations de transparence applicables à la publicité en ligne, le registre publicitaire prévu par le règlement du 13 mars 2024 devant permettre de compléter celui à mettre en œuvre dans le cadre de l'article 39 du RSN ;
  - les modalités d'application des avis de transparence prévus par le règlement du 13 mars 2024, qui doivent permettre d'assurer la traçabilité des flux entre tous les acteurs de la chaine de valeur publicitaire, tout particulièrement entre les plateformes et les utilisateurs susceptibles de bénéficier d'une large audience, comme les influenceurs diffusant des messages de publicité politique.

Sous-recommandation 1.2 : Adopter des lignes directrices en lien avec la manipulation de l'information, en application de l'article 35 du RSN

Depuis 2018, la lutte contre la manipulation de l'information en ligne s'est largement appuyée sur le code de bonnes pratiques contre la désinformation, initialement fondé sur une démarche volontaire d'autorégulation de la part de certaines plateformes. Renforcé en 2022 et intégré au RSN en 2025 en tant que code de conduite au sens de l'article 45, cet instrument est un outil de corégulation associant les plateformes signataires, les autorités de régulation, les acteurs industriels et la société civile. Les rapports publiés par les signataires du code, y compris ceux couvrant la période

[97] V. communiqué de presse précité du 5 décembre 2025. Dans le cadre d'une procédure d'enquête ouverte le 19 février 2024, la Commission a accepté des engagements contraignants pris par TikTok pour se conformer à ses obligations relatives à son registre publicitaire (article 39 RSN).
[98] https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=OJ:L_202400900

postérieure à son intégration au règlement, sont accessibles via le site suivant : https://disinfocode.eu/.

Si ce cadre a permis d'engager un dialogue structuré et de développer des bonnes pratiques, le durcissement du contexte géopolitique, l'accroissement de la menace informationnelle et l'expérience acquise dans les échanges avec les plateformes invitent aujourd'hui à évoluer vers une approche plus structurée et prescriptive à leur égard.

Dans ce contexte, dans le prolongement de la Stratégie nationale de lutte contre les manipulations de l'information d'origine étrangère (2026-2030), **l'Arcom recommande que des lignes directrices soient élaborées par la Commission européenne, en coopération avec le Comité européen des services numériques (*DSA Board*), tel que prévu par l'article 35 du RSN. Celles-ci permettraient de consolider le cadre issu du code et préciser les attentes des autorités de régulation européennes et nationales à l'égard des grandes plateformes en matière d'atténuation des risques systémiques liés à la manipulation de l'information en ligne**, tout en préservant, dans l'esprit du RSN, leur marge de manœuvre pour privilégier des mesures tenant compte des particularités de leurs services.

Ces lignes directrices pourraient notamment porter sur plusieurs axes prioritaires au vu des observations faites dans le présent bilan :

- **actualiser la liste de techniques, tactiques et procédures (*TTPs*)** utilisées dans le cadre d'opérations de manipulation de l'information, celle développée dans le cadre du code de conduite étant devenue en partie obsolète. En effet, les méthodologies employées par les acteurs malveillants évoluent rapidement, et l'émergence de nouvelles technologies, comme l'IA générative (utilisée par exemple par les *agents conversationnels*), modifie en permanence les vecteurs de manipulation[99]. À cet égard, l'Arcom appelle de ses vœux la **clarification du statut juridique des agents conversationnels d'IA générative intégrés dans les plateformes en ligne**, afin d'établir dans quelle mesure ces services relèvent du RSN et du régime de responsabilité qui en découle, et souligne l'importance de renforcer la vigilance des autorités de régulation à l'égard de la conception et du fonctionnement algorithmique de ces nouveaux services numériques, ainsi que de l'utilisation qui en est faite ;

- **rendre pleinement effective l'interdiction des faux comptes** qui est mentionnée dans les conditions d'utilisation des plateformes et que ces dernières doivent supprimer en application du RSN (article 35)[100], et des comportements inauthentiques, par exemple en homogénéisant et en détaillant les critères d'évaluation des risques systémiques en lien avec ces phénomènes. En effet, les plateformes publient aujourd'hui des informations lacunaires et difficilement comparables en raison de méthodologies, d'indicateurs de suivi et de périmètres

[99] Par ailleurs, certaines pratiques de manipulation, déjà interdites par les conditions d'utilisation de certaines grandes plateformes, ne sont actuellement pas couvertes par la liste des *TTPs* du code de conduite. C'est notamment le cas des techniques propres aux moteurs de recherche ou des nouvelles pratiques comme l'utilisation d'émojis ou d'obscurcissement de texte pour contourner les dispositifs de détection.

[100] Les grandes plateformes *« devraient, en particulier, examiner la manière dont la conception et le fonctionnement de leurs services, ainsi que l'utilisation et la manipulation intentionnelles et, souvent, coordonnées de leurs services, ou la violation systémique de leurs conditions d'utilisation, contribuent à ces risques. Ces risques peuvent résulter, par exemple, de l'utilisation non authentique du service, telle que la création de faux comptes, l'utilisation de robots ou l'utilisation trompeuse d'un service, et d'autres comportements automatisés ou partiellement automatisés, susceptibles de conduire à la diffusion rapide et généralisée au public d'informations qui constituent un contenu illicite ou qui sont incompatibles avec les conditions générales d'une plateforme en ligne ou d'un moteur de recherche en ligne et qui contribuent à des campagnes de désinformation* » (cons. 84 du RSN).

d'analyse hétérogènes (v. plus haut)[101]. Seules des données plus précises et complètes permettraient de mener des analyses comparatives utiles ;

- **prescrire une transparence accrue sur la conception et le fonctionnement des systèmes de recommandation (y compris ceux des IA génératives)**, qui ne font l'objet d'aucune précision dans les rapports publiés par les grandes plateformes. Cette transparence revêt une importance encore plus aigüe lors de certaines périodes sensibles pour la vie démocratique, comme l'organisation de scrutins électoraux, afin de mieux éclairer les dynamiques de viralité des contenus. Des mesures complémentaires pourraient aussi être envisagées, comme un mécanisme de « *période de gel* » (ou *standstill*) qui permettrait, en tenant compte des circonstances particulières de chaque élection, de limiter ou de suspendre temporairement certaines modifications algorithmiques pendant ces périodes ;

- **prescrire une maitrise renforcée des risques liés à la monétisation abusive de la désinformation** sur les plateformes en ligne et à leur modèle économique[102], notamment à travers :
  - les mécanismes de monétisation susceptibles d'être exploités à des fins de manipulation de l'information[103] ;
  - les fraudes, les escroqueries ou les redirections vers des liens trompeurs ;
  - la priorisation algorithmique de certains contenus, susceptible de générer des taux d'engagement importants (comme les contenus polarisants ou sensationnalistes) ;
  - la monétisation du temps passé par les utilisateurs sur les plateformes ;
  - les programmes de financement et de répartition de la valeur entre les créateurs de contenus comme les influenceurs et les plateformes. À cet égard, dans un avis du 18 février 2026, l'Autorité de la concurrence a appelé les plateformes à davantage de transparence dans leurs relations commerciales avec les créateurs de contenus[104] ;
  - le partage de revenus publicitaires avec des utilisateurs peu transparents ;
  - les dispositifs de rémunération indexés sur la performance et les techniques de microciblage.

L'Arcom relève que, sans préjudice de l'effet des lignes directrices qu'elle appelle de ses vœux, le risque d'explosion de la désinformation à but lucratif est intimement lié au modèle économique publicitaire de certaines grandes plateformes et à l'exemption conditionnelle de responsabilité dont elles disposent. **L'Arcom invite donc les pouvoirs publics à s'interroger sur la pertinence du maintien de cette exemption conditionnelle de responsabilité à raison des contenus hébergés par les plateformes, en ce qu'elle s'applique aux publicités pour le compte de tiers diffusées à titre onéreux sur leurs services.**

À cet égard, certaines plateformes mettent d'ailleurs en avant, dans leurs conditions d'utilisation, qu'elles exercent un contrôle poussé, en amont de la diffusion de publicités,

[101] L'homogénéité des données fournies par les plateformes demeure un enjeu structurant pour disposer d'informations plus cohérentes, exploitables et comparables pour les autorités publiques, les chercheurs et la société civile.
[102] Cet axe s'inscrit en cohérence avec la Stratégie nationale de lutte contre les manipulations de l'information 2026-2030 et le code de conduite contre la désinformation, lequel comporte un chapitre sur le contrôle de la publicité en ligne, composé d'engagements en faveur de la démonétisation de la désinformation.
[103] Par exemple, en prescrivant une authentification renforcée des utilisateurs ayant accès aux mécanismes de monétisation des plateformes en ligne.
[104] Avis 26-A-02 du 18 février 2026 relatif au fonctionnement de la concurrence dans le secteur de la création de contenu vidéo en ligne en France.

susceptible de leur conférer une connaissance, voire un contrôle, des données relatives aux annonces.

Tout en renforçant le cadre de supervision des risques systémiques sur les grandes plateformes, il apparaît essentiel de **préserver les structures de coopération mises en place dans le cadre du code de conduite contre la désinformation, qui constituent aujourd'hui des espaces d'échange utiles entre la Commission européenne, les plateformes signataires, les autorités de régulation nationales, les acteurs industriels et la société civile**. Par exemple, la pérennisation - dans un cadre plus structuré - de mécanismes de dialogue et de coopération comme le système de réponse rapide en contexte électoral (*Rapid Response System*), apparait importante.

Recommandation 2 : Mettre en place des mécanismes de financement pérennes pour soutenir les travaux de la recherche et des vérificateurs de faits

Acteurs visés : pouvoirs publics nationaux, Commission européenne

Les vérificateurs de faits et les chercheurs participent à la préservation de l'intégrité de l'espace informationnel en ligne. Ils permettent d'identifier rapidement les contenus faux ou trompeurs, d'évaluer l'efficacité des mesures mises en œuvre par les grandes plateformes à travers une démarche scientifique, de documenter ces phénomènes ou encore de sensibiliser les utilisateurs sur les menaces émergentes. Leurs analyses sont également utiles aux coordinateurs pour les services numériques, tels que l'Arcom, chargés de superviser l'application du RSN dans leur État membre.

En France, le programme De Facto[105], lancé et financé par la Commission européenne, qui regroupe des chercheurs, des journalistes, et professionnels de l'éducation aux médias et à l'information, a pour ambition de promouvoir la qualité de l'information, la diversité du débat public et la régulation des plateformes en ligne.

Aujourd'hui, comme en attestent les appels à soutien des vérificateurs de faits, ces acteurs rencontrent d'importantes difficultés pour obtenir des financements rapides et durables, tant au niveau national qu'européen. Ces difficultés, qui menacent la continuité et l'efficacité de leur travail, sont accentuées par l'incertitude régnant sur la pérennité des programmes de soutien aux vérificateurs de faits de certaines grandes plateformes.

À l'approche des élections présidentielles de 2027, un soutien public stable aux chercheurs et aux vérificateurs de faits est indispensable pour leur permettre de continuer à documenter les phénomènes observés sur les plateformes.

[105] De Facto est la branche française du projet européen EDMO (Observatoire européen des médias numériques) - https://defacto-observatoire.fr/

Recommandation 3 : Renforcer la mobilisation collective, en créant un Observatoire sur la lutte contre les manipulations de l'information

Acteurs visés : pouvoirs publics nationaux, plateformes, société civile

Le renforcement des capacités nationales d'analyse et de suivi des phénomènes de manipulation de l'information, qui ne limitent pas aux cas d'ingérence étrangère, pourrait également se matérialiser par la création d'un dispositif structuré d'échange entre plateformes, autorités et acteurs spécialisés (société civile, chercheurs, acteurs industriels, etc.), comme suggéré par la Stratégie nationale de lutte contre les manipulations de l'information d'origine étrangère (2026-2030).

L'Arcom considère comme cette dernière que ce dispositif pourrait prendre la forme d'un « Observatoire sur la lutte contre les manipulations de l'information », en s'inspirant *mutatis mutandis* de l'Observatoire de la haine en ligne, placé auprès de l'Arcom. En tant que coordinateur pour les services numériques pour la France, et forte des coopérations déjà nouées avec les plateformes, la société civile et le monde académique, l'Arcom pourrait assurer la structuration, le pilotage et l'animation de cet Observatoire.

Cet espace d'échange soutiendrait le développement d'une expertise partagée en matière de manipulation de l'information en ligne, en ce qu'il permettrait notamment un partage de connaissance et des échanges de bonnes pratiques (ex. identification de menaces émergentes, évolution de phénomènes observées sur les plateformes, etc.), qui aurait *in fine* pour objectif d'alimenter les travaux nationaux et européens en la matière.

L'Arcom, en coopération avec le SGDSN, le Ministère de l'Europe et des Affaires étrangères, le Ministère de l'Intérieur, la Commission des sondages et la Commission européenne, anime d'ores et déjà une telle structure avant chaque élection depuis 2018. En plus des acteurs publics précités, elle rassemble les principales grandes plateformes et des acteurs de la société civile (notamment membres du projet De Facto).

L'institutionnalisation de ce cadre, sous la forme d'un « Observatoire sur la lutte contre les manipulations de l'information », son élargissement, notamment à la communauté de la recherche, et son activation hors période électorale permettrait d'inscrire l'effort collectif dans la durée et de mobiliser les acteurs sur des problématiques non-électorales prioritaires comme la lutte contre la désinformation en santé ou en matière environnementale.

Recommandation 4 : Renforcer et élargir les actions d'éducation aux médias, à l'information et à la citoyenneté numérique à de nouveaux réseaux et publics

Acteurs visés : pouvoirs publics nationaux et acteurs de l'éducation aux médias, à l'information et à la citoyenneté numérique

Les phénomènes de manipulation de l'information prennent aujourd'hui des formes multiples et peuvent toucher des publics variés.

Dans ce contexte, l'Arcom recommande aux acteurs de l'éducation aux médias, à l'information et à la citoyenneté numérique d'élargir le périmètre de leurs actions à des publics situés en dehors du cadre scolaire. Si les dispositifs portés par l'Éducation nationale et le Centre pour l'éducation aux médias et à l'information (CLEMI) jouent un rôle structurant dans la sensibilisation des jeunes publics, les enjeux informationnels nécessitent de développer des initiatives complémentaires permettant de toucher d'autres publics, tels que les adultes et notamment les parents ainsi que les séniors, en s'appuyant sur de nouveaux relais comme les collectivités territoriales ou les structures associatives locales.

L'étude publiée par l'Arcom le 25 septembre 2025, intitulée « Mineurs en ligne : quels risques, quelles protections ? », est à ce titre particulièrement révélatrice du rôle déterminant des aînés dans l'accompagnement des usages numériques des plus jeunes. Aussi, certains publics plus isolés, en raison de leur potentiel accès limité à des sources d'information fiables ou de leur faible exposition aux dispositifs de sensibilisation, pourraient bénéficier de ces initiatives locales adaptées à leurs besoins et au contexte dans lequel ils évoluent.

Dans le prolongement de la Stratégie nationale de lutte contre les manipulations de l'information d'origine étrangère (2026-2030)[106], il est particulièrement important de ne pas limiter les actions d'éducation aux médias, à l'information et à la citoyenneté numérique aux seuls cas d'ingérences numériques étrangères et de prendre en considération les phénomènes de manipulation de l'information domestiques, qui peuvent tout autant altérer la qualité du débat public, influencer la perception des citoyens et fragiliser la confiance dans l'information.

En élargissant l'éducation aux médias, à l'information et à la citoyenneté numérique à de nouveaux publics et en intégrant l'ensemble des menaces liées à la manipulation de l'information, il s'agit ainsi de renforcer la résilience informationnelle de l'ensemble de la société et de favoriser le développement d'un esprit critique chez les citoyens, leur permettant d'analyser et de comprendre les dynamiques à l'œuvre derrière les contenus circulant sur les plateformes en ligne.

[106] Objectifs stratégiques 1, 2 et 4.

Annexe 1 – CGU par grande plateforme en lien avec les *TTPs*

Avertissement

La présente annexe recense les politiques des grandes plateformes pertinentes au regard des *TTPs*. Les informations qui y figurent sont à jour en mars 2026, sans préjudice de leurs évolutions ultérieures.

Meta

1. Les médias synthétiques et manipulés[107]

Meta prend en compte les critères suivants pour identifier les médias synthétiques et manipulés :

- un contenu créé ou retouché numériquement ;
- un contenu susceptible d'induire en erreur.

Meta peut placer une étiquette informative sur le média synthétique ou manipulé, lorsqu'il s'agit d'une image, d'une vidéo ou d'un contenu audio qui semble réaliste, créé ou retouché numériquement et qu'il implique un risque particulièrement élevé de tromper considérablement le public sur un sujet d'intérêt général. Meta peut aller jusqu'à rejeter le contenu s'il s'agit d'une publicité.

2. Fraude et usurpation d'identité[108] [109]

L'usurpation d'identité est interdite sur les services de Meta, qui a également mis en place une politique spécifiquement intitulée « *fraude, arnaque et pratiques trompeuses* », qui englobe une série de contenus interdits dans plusieurs domaines (ex. domaine financier, identité non authentique, emploi, jeux etc.).

Concernant les identités non authentiques, Meta précise qu'il est interdit, sur ses services, de diffuser des contenus qui essayent d'arnaquer ou d'escroquer les utilisateurs en mentant sur l'identité de l'auteur de la publication ou la nature d'une demande (ex. fraudes liées aux associations caritatives, à une entreprise établie, etc.).

3. Comportements inauthentiques et engagements artificiels[110]

Meta précise que « *le comportement non authentique fait référence à une variété de formes complexes de tromperie dont se rend coupable un réseau d'éléments non authentiques contrôlés par le ou les mêmes individus dans le but de tromper Meta ou [sa] communauté, ou d'échapper à l'obligation de respecter les Standards de la Communauté* ».

Sont interdits au titre de cette politique :

- la création ou l'utilisation d'éléments non authentiques (comptes, pages, groupes, etc.) dans le but de :
  - tromper Meta ou les utilisateurs :

[107] https://transparency.meta.com/policies/community-standards/misinformation (dernier paragraphe)
[108] https://transparency.meta.com/policies/community-standards/fraud-and-scams
[109] https://transparency.meta.com/policies/community-standards/inauthentic-behavior
[110] https://transparency.meta.com/policies/community-standards/inauthentic-behavior

  - sur l'identité, l'objectif, ou l'origine d'une audience ou de l'entité représentée ;
  - sur la popularité du contenu ou d'éléments sur ses services ;
  - sur le réseau de propriété ou de contrôle d'un élément actif Meta.
- contourner des dispositions des standards de la communauté ;
- utiliser de manière détournée les systèmes de signalements Meta pour harceler, intimider ou réduire au silence d'autres personnes.

Meta demande du contexte et/ou des informations supplémentaires pour assurer le respect de ses standards de la communauté pour les cas suivants :

- les entités qui adoptent ou prétendent adopter un comportement non authentique coordonné, défini comme des formes particulièrement sophistiquées de comportement non authentique, où les fausses identités sont primordiales pour l'opération. Dans ce cadre, les acteurs malveillants utilisent des méthodes malveillantes pour échapper à la détection ou paraître authentique ;
- les entités qui se livrent ou prétendent se livrer à une ingérence étrangère, définie comme un comportement non authentique coordonné pour lequel les acteurs du réseau ne sont pas situés dans le même pays que l'audience ciblée par l'opération.

4. Contenus indésirables (spams)[111]

La politique dédiée au contenu indésirable de Meta interdit :

- toute technique abusive (ex. pop-ups non pertinents, détournements de clics, publication de liens fallacieux ou mensongers[112]) destinés à attirer les utilisateurs vers des fonctionnalités ou des codes trompeurs, ou à se faire passer pour un domaine de confiance ;
- la publication, le partage manuel ou automatique de contenus à des fréquences très élevées ;
- la création de comptes, groupes, pages, événements ou autres éléments, à des fréquences très élevées ;
- le fait de tenter de vendre, d'acheter ou d'échanger des actifs de la plateforme comme des comptes, des groupes, des pages ou d'y parvenir ;
- le fait de tenter de vendre, d'acheter ou d'échanger des privilèges du site comme des rôles d'administrateur ou de modérateur ou des autorisations de publier dans des espaces spécifiques, ou d'y parvenir ;
- le fait de tenter de vendre, d'acheter ou d'échanger des interactions, comme les likes, les partages, les vues, les abonnements, les clics ou l'utilisation de certains hashtags, ou d'y parvenir ;
  - ceci inclut la proposition de concours permettant de gagner de l'argent ou un équivalent en échange d'interactions (ex. « *aimez ma page et tentez de remporter 500€* ») ou de proposer de fournir quelque chose ayant une valeur monétaire en échange d'interactions (ex. « *si vous aimez ma page, je vous offre un iPhone* ») ;
- d'exiger ou de réclamer que les utilisateurs interagissent avec du contenu (ex. aimer ou partager) avant de pouvoir voir ou interagir avec le contenu promis.

[111] https://transparency.meta.com/policies/community-standards/spam/
[112] Par exemple, via la technique du masquage (*cloaking*), utilisée pour contourner les processus d'évaluation de Facebook et Instagram et afficher du contenu qui enfreint les normes communautaires et les politiques publicitaires de Meta.

1. Les médias synthétiques et manipulés[113]

X définit le média manipulé comme étant un « *média partagé comme authentique, mais qui est considérablement et de manière trompeuse, altéré, manipulé ou fabriqué d'une manière à en changer fondamentalement le sens et à entraîner une confusion généralisée sur des questions publiques, à avoir un impact sur la sécurité publique ou causer de graves dommages* ».

Cela comprend :

- les médias qui ont été substantiellement modifiés ou post-traités d'une manière qui altère fondamentalement leur composition, leur séquence, leur rythme ou leur cadrage, et qui en déforme le sens ;
- les médias comportant des informations visuelles ou auditives (ex. un doublage audio ou des sous-titres modifiés) qui ont été ajoutées, modifiées ou supprimées et qui changent fondamentalement la compréhension, la signification ou le contexte des médias ;
- médias créés, édités ou post-traités avec des améliorations ou l'utilisation de filtres qui modifient fondamentalement la compréhension, la signification ou le contexte du contenu ;
- médias représentant une personne réelle qui a été fabriquée ou simulée, notamment par l'utilisation d'algorithmes ou une intelligence artificielle.

X définit ensuite le média hors contexte, que la plateforme considère comme des « *médias qui ne sont pas manipulés, mais qui sont partagés de manière trompeuse ou hors contexte, ou dans l'intention de tromper le public sur la nature ou l'origine du contenu, et qui peuvent entraîner une confusion généralisée sur des questions publiques, avoir un impact sur la sécurité publique ou causer des dommages graves* ».

Cela comprend :

- les médias présentés avec un contexte faux ou trompeur concernant la source, le lieu, l'heure ou l'authenticité du média ;
- les médias présentant un contexte faux ou trompeur concernant l'identité des individus ou des entités représentés visuellement ;
- les médias présentant des déclarations ou des citations erronées ou qui présentent des affirmations fabriquées à partir de faits réels.

2. Fraude et usurpation d'identité[114]

X précise que l'usurpation d'identité constitue une infraction à ses règles. La plateforme mentionne que, bien qu'il ne soit pas obligatoire d'afficher sa véritable identité sur son profil, le compte ne doit pas utiliser des informations qui usurpent l'identité d'autrui. X mentionne toutefois autoriser les comptes de parodie, de commentaires et de fans qui respectent ses CGU, uniquement si leur but est de discuter, de satiriser ou de partager des informations.

Les actions de fraude et d'escroquerie sont également prohibées par la plateforme, qui cite certaines tactiques trompeuses qu'elle interdit expressément :

[113] https://help.x.com/en/rules-and-policies/authenticity (partie sur les médias synthétiques et manipulés).
[114] https://help.x.com/en/rules-and-policies/authenticity (parties sur l'usurpation d'identité ; les escroqueries).

- l'ingénierie sociale, notamment les systèmes d'établissement de relations de confiance ;
- les systèmes d'offre d'argent en échange d'un premier versement ;
- les rabais frauduleux ;
- l'hameçonnage.

3. Comportements inauthentiques et engagements artificiels[115]

X interdit d'adopter des comportements visant à manipuler X ou à impacter de manière artificielle la façon dont les contenus sont amplifiés.

Au sujet des **comptes inauthentiques**, X n'autorise pas la création, l'exploitation ou l'enregistrement massive de comptes qui ne sont pas légitimes, authentiques et transparents quant à leur source, leur identité et leur popularité. Cela inclut :

- l'automatisation non autorisée ;
- les faux profils ;
- l'usurpation d'identité.

En ce qui concerne la **coordination de comptes inauthentiques multiples**, X interdit l'exploitation de plusieurs comptes qui interagissent avec le même contenu ou un contenu sensiblement similaire afin de gonfler ou de manipuler l'importance du contenu et/ou des comptes (influence artificielle), par exemple en créant plusieurs comptes pour :

- stimuler les sujets tendance ou les hashtags ;
- interagir avec les mêmes publications, comptes et sondages ;
- abuser de la fonction de mention/réponse ;
- amplifier l'un de ses propres comptes en utilisant abusivement les fonctionnalités d'engagement.

X interdit également d'exploiter plusieurs comptes qui publient du contenu sensiblement similaire ou identique les uns aux autres, par exemple en publiant de manière croisée :

- des contenus sur plusieurs comptes ;
- des contenus similaires ou dupliqués sur les mêmes sujets « *tendance* » ou hashtags.

L'utilisation de solutions de contournement pour dépasser les limites techniques de création de compte est prohibée. À noter qu'X permet aux utilisateurs de créer et/ou d'exploiter jusqu'à 10 comptes à des fins différentes et non duplicatives.

4. Contenus indésirables (spams)[116]

X interdit sur son service :

- **les spams d'engagements**, qui comprennent :
  - la coordination de l'échange d'engagement dans toutes les fonctionnalités d'X, telles que les mentions « *J'aime* », les sondages, les réponses, les republications, les listes, les vues et les abonnements ;
  - la coordination et/ou la rémunération d'autres personnes pour effectuer une inflation des indicateurs de comptes dans toutes les fonctionnalités susmentionnées ;
  - le fait de se livrer à des abonnements/désabonnements en masse (« *follow churn* »), c'est-à-dire de suivre puis de se désabonner immédiatement d'un

[115] https://help.x.com/en/rules-and-policies/authenticity (partie sur les comptes et comportements inauthentiques ; les comptes multiples et coordonnés).
[116] https://help.x.com/en/rules-and-policies/authenticity (partie sur les spams).

grand nombre de comptes dans le but d'augmenter son propre nombre d'abonnés ;
- le fait de se livrer à un suivi indiscriminé, autrement dit de suivre et/ou ne plus suivre un grand nombre de comptes sans rapport entre eux sur une courte période, notamment par des moyens automatisés ;
- le fait d'interagir avec les publications de manière agressive ou en utilisant l'automatisation pour générer du trafic ou attirer l'attention vers des comptes, des sites web, des produits, des services ou des initiatives ;
- le fait de participer à l'ajout agressif d'utilisateurs à des listes pour leur diffuser du contenu ;
- la duplication des abonnés d'un autre compte, notamment en utilisant des moyens automatisés ;
- l'utilisation ou la promotion des services tiers pour effectuer l'une des transactions susmentionnées ;
- l'interaction avec les fonctionnalités ou formulaires de signalement d'X, de façon automatisée ou non, afin de soumettre des rapports en double ou en grand nombre, y compris en signalant à plusieurs reprises les mêmes comptes, ou de coordonner ou encourager d'autres personnes à utiliser abusivement lesdites fonctionnalités.

- **les spams de contenu**, qui comprennent :
  - les envois massifs, agressifs et volumineux de réponses, de mentions ou de messages directs non sollicités ;
  - l'utilisation d'hashtags tendance ou populaires dans le but de subvertir, de manipuler une conversation ou de générer du trafic ou de l'attention vers des comptes, des sites web, des produits, des services ou des initiatives ;
  - la publication de contenu avec des hashtags excessifs et sans rapport ;
  - la publication ou l'envoi de manière répétée de messages directs composés de liens partagés sans commentaire ;
  - la promotion d'un contenu en répondant à celui-ci avec du contenu qui n'est pas pertinent par rapport au sujet du message d'origine ;
  - la publication et la suppression du même contenu à plusieurs reprises ;
  - la publication de manière répétée de messages identiques ou presque identiques de façon dupliquée (« *copypasta* »), ou l'envoi de messages directs identiques ;
  - la modification de manière trompeuse d'un contenu ayant déjà un engagement existant, pour amplifier un contenu sensiblement différent (ex. modifier une publication de « *qu'est-ce-qui est le mieux entre les crêpes et les gaufres ?* » à « *des milliers de personnes font confiance à mon service. Aimez ma publication et abonnez-vous à ma chaîne* ») ;
  - la modification de liens URL de sorte que la page de destination finale ait considérablement changé, soit en termes de contenu, soit en termes d'emplacement web.

1. Les médias synthétiques et manipulés[117]

YouTube distingue les contenus « *manipulés* » des contenus « *attribués à tort* ».

Les **contenus « *manipulés* »** sont des contenus qui ont été techniquement manipulés ou falsifiés de manière à induire les utilisateurs en erreur et qui peuvent présenter un risque important de préjudice majeur.

Exemples :
- *sous-titres des vidéos traduits de manière incorrecte, qui attisent les tensions géopolitiques et créent un risque sérieux de préjudice grave ;*
- *vidéos qui ont été modifiées à l'aide de moyens techniques (généralement au-delà de simples extraits sortis de leur contexte) pour faire croire qu'un responsable gouvernemental est décédé ou pour inventer des événements présentant un risque de préjudice important.*

Les **contenus « *attribués à tort* »** sont des contenus présentant un risque de préjudice majeur en prétendant faussement qu'une ancienne vidéo d'un événement passé représente un événement actuel.

Exemple : contenu présenté de manière inexacte comme documentant des violations des droits de l'homme dans un lieu précis, alors qu'il s'agit en réalité d'un contenu provenant d'un autre lieu ou événement.

2. Fraude et usurpation d'identité[118]

YouTube fournit une explication des pratiques d'« *usurpation d'une chaîne* » et d'« *usurpation d'identité d'une personne* », que la plateforme interdit :

- **L'usurpation d'une chaîne** se produit lorsqu'une chaîne reprend le profil, l'arrière-plan ou l'apparence générale d'une autre chaîne dans le but de se faire passer pour son propriétaire. Cela ne signifie pas nécessairement que la chaîne est rigoureusement identique, mais que son intention de copier une autre chaîne est claire. Par exemple :
  - les chaînes avec le même nom ou identifiant, et la même image qu'une autre chaîne, la seule différence étant un espace ajouté dans le nom ou un zéro remplaçant la lettre O ;
  - les chaînes utilisant le véritable nom d'utilisateur, les images, la marque, le logo ou d'autres informations personnelles d'une tierce personne dans le but de tromper le public et de se faire passer pour cette personne ;
  - les chaînes configurées avec le nom et l'image d'une personne pour prétendre que cette personne publie du contenu sur la chaîne ;
  - les chaînes configurées avec le nom et l'image d'une personne, qui publient des commentaires sur d'autres chaînes en se faisant passer pour cette personne ;

[117] https://support.google.com/youtube/answer/10834785 (règlement communautaire).
[118] https://support.google.com/youtube/answer/2801947 (règlement communautaire).

- les chaînes prétendant être un « compte de fan » dans leur description, mais qui ne l'indiquent pas clairement dans leur nom ou leur identifiant, ou qui se font passer pour une autre chaîne et remettent en ligne ses contenus ;
- les chaînes qui usurpent l'identité d'une chaîne d'information.

- **L'usurpation d'identité d'une personne** se matérialise par un contenu conçu pour faire croire qu'il a été publié par une autre personne.

Par ailleurs, YouTube mentionne que les escroqueries et autres pratiques trompeuses visant à exploiter la communauté sont interdites sur la plateforme. Une liste non exhaustive de types d'escroquerie est présentée par YouTube, notamment :
- les promesses exagérées, comme prétendre que les spectateurs peuvent s'enrichir rapidement ou qu'un traitement miraculeux peut guérir des maladies chroniques ;
- la mise en ligne des vidéos frauduleuses (ex. « *Grâce à ce programme, vous allez gagner 50 000€ en une journée !* ») ;
- la création de comptes dédiés à des programmes proposant d'offrir des récompenses en espèces, etc.

3. Comportements inauthentiques et engagements artificiels[119]

YouTube n'autorise aucune pratique visant à augmenter artificiellement le nombre de vues, de *likes*, de commentaires ou d'autres métriques par l'utilisation de systèmes automatisés ou par la diffusion auprès d'utilisateurs non avertis. Les contenus dont l'unique but est de provoquer l'engagement des spectateurs sont également interdits.
Cela concerne notamment :
- le contenu qui fait la promotion ou renvoie vers des services tiers permettant d'augmenter artificiellement les métriques telles que le nombre de vues, de *likes* ou d'abonnés ;
- le contenu qui fait la promotion ou renvoie vers des sites web ou des services tiers proposant de tromper le système afin de fausser le nombre de vues ou d'abonnés ;
- le contenu dans lequel un utilisateur propose à d'autres créateurs de s'abonner à leur chaîne uniquement si ceux-ci s'abonnent à la sienne en retour (*sub4sub*)[120] ;
- le contenu dans lequel un créateur achète des vues auprès d'un service tiers dans le but de promouvoir ce service.

Ces règles s'appliquent aux vidéos, aux descriptions de vidéos, aux commentaires, aux diffusions en direct et à tout autre produit ou fonctionnalité de YouTube (liste non exhaustive).

4. Contenus indésirables (spams)[121]

YouTube dédie une politique spécifique concernant « *le spam, les pratiques trompeuses et les escroqueries* », et distingue également plusieurs catégories de spams interdits, suivant leur emplacement sur l'interface ou leur but :

- **les spams vidéo** : contenu publié de manière excessive, répétitif ou non ciblé, et qui :

[119] https://support.google.com/youtube/answer/3399767 (règlement communautaire sur l'engagement artificiel).
[120] À noter qu'un créateur a le droit d'encourager ses spectateurs à s'abonner, à aimer ses contenus ou à laisser un commentaire.
[121] https://support.google.com/youtube/answer/2801973 (règlement communautaire concernant le spam, les pratiques trompeuses et les escroqueries).

- laisse croire aux utilisateurs qu'ils vont voir des contenus, mais les redirige en dehors de YouTube ;
- génère des clics, des vues ou du trafic en dehors de YouTube, en promettant un gain d'argent rapide aux utilisateurs ;
- redirige les utilisateurs vers des sites qui propagent des logiciels dangereux, tentent de collecter des informations personnelles ou engendrent un impact négatif.

Exemples :
- *publier le même contenu à plusieurs reprises sur une ou plusieurs chaînes ;*
- *mettre en ligne massivement un contenu détourné d'autres créateurs ;*
- *publier du contenu généré automatiquement par des ordinateurs afin de le mettre en ligne sans se préoccuper de la qualité ou de l'expérience des utilisateurs ;*
- *publier en masse du contenu affilié à un compte dédié, etc.*

- **les spams dans les commentaires** : commentaires dont l'unique objectif est de collecter des informations personnelles sur les utilisateurs, de les conduire à quitter YouTube à leur insu ou d'adopter l'un des comportements définis dans cette politique.

Exemples :
- *les commentaires répétitifs et non ciblées ;*
- *les enquêtes ou offres faisant la promotion de systèmes pyramidaux ;*
- *les commentaires prétendant offrir des contenus vidéos dans leur intégralité (ex. films et séries) ;*
- *les liens vers des sites d'hameçonnage ou contenant des logiciels malveillants, etc.*

- **les spams incitatifs** : contenu proposant à la vente des métriques d'engagement telles que des vues, des mentions « *J'aime* », des commentaires ou tout autre indicateur disponible sur YouTube. Il peut également s'agir de contenus dans lesquels un créateur propose à un autre créateur de s'abonner à sa chaîne uniquement à condition qu'il s'abonne aussi en retour (*sub4sub*).

- **les métadonnées ou miniatures trompeuses** : utilisation du titre, des miniatures ou de la description pour tromper les utilisateurs sur la nature du contenu ;

- **les contenus tiers** : diffusions en direct incluant des contenus tiers non autorisés, et qui ne sont pas supprimées malgré l'envoi de plusieurs mises en garde concernant une possible utilisation abusive. Le propriétaire d'une chaîne doit surveiller activement ses diffusions en direct et corriger tout problème dans les meilleurs délais.

1. Les médias synthétiques et manipulés[122]

TikTok n'autorise pas le matériel visuel ou audio qui a été édité, assemblé, combiné, d'une manière qui pourrait induire un utilisateur en erreur sur des événements du monde réel.
À ce titre, sont par exemple interdits :

- les images créées par l'IA pour intimider ou harceler ;
- les contenus générés par l'IA ou modifiés de manière significative qui induisent en erreur sur un sujet d'importance publique[123].

TikTok peut rendre inéligible au fil « *For You* » tout contenu d'apparence réaliste dont il n'a pas encore été confirmé qu'il est du contenu généré par l'IA ou modifié de manière significative, mais qui présente des questions d'importance publique d'une manière qui pourrait conduire à une mauvaise interprétation ou nuire à des personnalités privées.

Exemple de contenus qui doivent être présentés comme tels ou étiquetés : contenus présentant des images, des vidéos ou de l'audio entièrement ou considérablement modifiés ou générés par l'IA.

Par « ***considérablement modifié par l'IA*** », TikTok entend un contenu modifié au-delà des simples ajustements ou améliorations mineures. Cela inclut l'utilisation d'images ou de vidéos réelles comme source, mais modifiés substantiellement par l'IA, par exemple :

- montrer que le sujet principal fait quelque chose qu'il n'a pas réellement fait ;
- faire dire au sujet principal quelque chose qu'il n'a pas réellement dit, en utilisant le clonage vocal de l'IA ;
- modifier l'apparence ou la voix du sujet principal à tel point qu'il n'est plus reconnaissable ou visible (comme l'échange de visage par l'IA). Cela comprend :
  - recadrer ou couper des phrases pour en changer le sens ;
  - réorganiser ou combiner des extraits ;
  - changer la vitesse ou ajouter/supprimer des parties audio/vidéos.

Par « ***simples ajustements ou améliorations mineures*** », TikTok entend par exemple le réglage de l'éclairage, de la luminosité ou de la saturation des couleurs, la suppression ou la modification de l'arrière-plan, la réduction du niveau du bruit.

Par ailleurs, tous les contenus présentant des images, des vidéos ou de l'audio entièrement ou considérablement modifiés ou générés par l'IA doivent être présentés comme tels ou étiquetés.

[122] https://www.tiktok.com/community-guidelines/en/integrity-authenticity#3 (partie sur les contenus multimédias édités et les contenus générés par l'IA).
[123] TikTok fournit une liste non exhaustive des « *sujets d'importance publique* » : contenu conçu pour donner l'impression qu'il provient d'une véritable source d'information ; événement de crise, catastrophe naturelle, conflit ; personnalité publique rabaissée, harcelée ou associée à un comportement criminel ; personnalité publique prenant position, faisant la promotion de produits ou commentant des problèmes publics qu'elle n'a pas réellement abordés ou encore une condamnation politique qui ne s'est jamais produite.

2. Fraude et usurpation d'identité[124]

TikTok interdit le fait de se faire passer pour quelqu'un d'autre sans indiquer clairement dans le nom d'affichage que le compte est un compte de fan ou parodique.

La plateforme prohibe également tout contenu qui promeut ou facilite les escroqueries, les fraudes ou les stratagèmes trompeurs.

3. Comportements inauthentiques et engagements artificiels[125]

TikTok n'autorise pas les comptes qui induisent en erreur ou qui tentent de manipuler la plateforme, ni le commerce de services qui stimulent artificiellement l'engagement ou trompent le système de recommandation. Cela inclut des comportements tels que les opérations d'influence, l'usurpation d'identité, le contenu indésirable, les faux avis et le partage de contenu piraté de manière nuisible. TikTok interdit strictement les outils d'automatisation, les scripts ou les astuces conçues pour contourner ses systèmes.

La plateforme définit les **opérations d'influence** comme des comportements coordonnés inauthentiques caractérisés par des réseaux de comptes qui collaborent pour tromper les utilisateurs ou les systèmes de TikTok et tenter d'influencer stratégiquement le débat public.
Au titre des opérations d'influence, sont interdits par les CGU de TikTok :

- le fait de tenter d'influencer de manière trompeuse les élections, les questions sociales, la politique ou les conflits armés ;
- les comptes coordonnés secrètement pour promouvoir un candidat ou un sujet politique ;
- la publication de contenu au nom d'entités étrangères (comme un gouvernement ou une armée), sans en faire mention ;
- le contenu indésirable, tel que l'achat ou la vente de followers ou d'engagements à des fins financières ;
- l'utilisation de bots ou de scripts pour écrire de faux avis ou commentaires, ou pour augmenter le nombre de « *J'aime* » ou les partages ;
- le contenu qui permet de négocier, commercialiser ou donner accès à des services qui augmentent artificiellement l'engagement, tels que : les followers ou les « *J'aime* », les faux avis ou l'utilisation de comptes d'IA ou de bots pour générer du trafic ;
- le partage de guides pratiques ou de conseils pour stimuler l'engagement de manière fausse ou trompeuse.

TikTok rend par ailleurs inéligible au fil « *For You* » toute action visant à inciter les utilisateurs à accroitre leur engagement, notamment via :

- les promesses de type *sub4sub* ;
- les fausses incitations pour recevoir des cadeaux ou augmenter le nombre d'abonnés ;
- les affirmations trompeuses destinées à augmenter les vues ou la popularité.

[124] https://www.tiktok.com/community-guidelines/en/integrity-authenticity#5 (mentionnées en marge de parties thématiques clairement identifiées comme les contenus générés par IA ou les comportements trompeurs et les faux engagements).
[125] https://www.tiktok.com/community-guidelines/en/integrity-authenticity#5 (partie sur les comportements trompeurs et les faux engagements).

4. Contenus indésirables (spams)[126]

Les politiques de TikTok interdisent les comportements trompeurs, inauthentiques et mensongers, afin de protéger l'intégrité, l'authenticité et la sécurité des utilisateurs. TikTok prévoit des politiques en la matière pour les contenus organiques et publicitaires. À ce titre, la plateforme interdit :

- **les pièges à clics** : contenus utilisant des éléments visuels ou des tactiques visant à tromper, induire en erreur ou manipuler les utilisateurs pour les inciter à interagir avec du contenu.

Exemples :
- *des éléments non cliquables ou trompeurs qui créent de fausses attentes (ex. faux bouton de lecture vidéo ou de fermeture d'annonce) ;*
- *le contenu dissimulé ou trompeur (ex. images suggestives floues ou bloquées ; contenu sexuel implicite ou de langage trompeur pour inciter les utilisateurs à cliquer, etc.).*

- **l'exploitation de comptes de spam,** en ce compris l'utilisation de l'automatisation pour gérer de nombreux comptes ou envoyer du contenu répétitif ;

- la publication d'une grande quantité de contenu non pertinent ;

- l'achat ou la vente d'abonnés ou d'engagements à des fins financières.

[126] https://www.tiktok.com/community-guidelines/fr/integrity-authenticity#5 (la partie sur les contenus indésirables est hébergée au sein de la thématique des comportements trompeurs et des faux engagements. Il est nécessaire de cliquer sur l'onglet « *informations supplémentaires* » pour voir dérouler un fil d'actions non autorisées, comme les contenus indésirables).

1. Les médias synthétiques et manipulés[127]

Microsoft interdit, pour l'ensemble de ses services, la création et la diffusion de contenus trompeurs générés par IA. Cela inclut les contenus audios, vidéos et les images générés par IA qui falsifient ou modifient de manière trompeuse l'apparence, la voix ou les actions d'autres personnes[128].
LinkedIn interdit les médias manipulés ou synthétiques trompeurs, qui ne divulguent pas clairement la nature fausse ou modifiée du contenu.

Exemples :
- *images ou vidéos trafiquées, qui déforment des événements réels et sont susceptibles de nuire au sujet, à d'autres individus ou groupes, ou à la société dans son ensemble ;*
- *contenu photoréaliste ou audio-réaliste qui représente quelqu'un disant quelque chose qu'il n'a pas dit ou faisant quelque chose qu'il n'a pas fait.*

2. Fraude et usurpation d'identité[129]

Microsoft interdit toute forme d'arnaque, de fraude, d'hameçonnage ou de pratiques trompeuses, y compris l'usurpation d'identité, sur ses plateformes et services.

Microsoft fait référence à tout acte ou omission intentionnelle visant à tromper autrui afin d'en tirer un avantage personnel ou financier. L'hameçonnage comprend notamment l'envoi de courriels ou d'autres communications électroniques dans le but d'inciter frauduleusement ou illégalement les destinataires à divulguer des informations personnelles ou sensibles.

Le fournisseur de services fournit une liste d'exemples d'escroquerie, de fraude et d'hameçonnage :
- fausses promesses aux utilisateurs, concernant une offre légitime ou pertinente, mais qui les redirige en réalité vers un site externe ;
- proposition de cadeaux, de programmes d'enrichissement rapide, d'intégration dans des systèmes pyramidaux ;
- vente d'indicateurs d'engagement, tels que les vues, les mentions « *J'aime* » ou les commentaires ;

[127] https://fr.linkedin.com/legal/professional-community-policies (paragraphe relatif au contenu faux ou trompeur des politiques de la communauté professionnelle) ou https://www.linkedin.com/help/linkedin/answer/a1340752/ (centre d'aide de LinkedIn) ; https://blogs.microsoft.com/on-the-issues/2024/02/13/generative-ai-content-abuse-online-safety/ (paragraphe sur la protection des services de Microsoft sur les contenus et comportements abusifs) ; https://www.microsoft.com/fr-fr/digitalsafety/policies (paragraphe sur le contenu généré par IA trompeur en matière électorale).
[128] À noter que Microsoft intègre désormais Copilot au sein de l'ensemble de ses produits, dont Microsoft Bing. Les conditions d'utilisation de Copilot, qui reprennent les politiques générales de Microsoft, s'appliquent donc aux expériences IA de Microsoft Bing.
[129] https://www.microsoft.com/fr-fr/digitalsafety/policies (paragraphe sur l'escroquerie, la fraude et l'hameçonnage, qui comprend l'interdiction de l'usurpation d'identité) ; https://fr.linkedin.com/legal/user-agreement? (partie 8.2.1 « *Règles* »).

toute tentative visant à tromper les utilisateurs afin de les inciter à visiter des sites web destinés à faciliter la propagation de logiciels malveillants ou espions, etc.

3. Comportements inauthentiques et engagements artificiels[130]

Au sein du *Microsoft Services Agreement*, qui régit notamment l'utilisation de Bing et de LinkedIn, il est stipulé que les activités frauduleuses, fausses ou trompeuses (ex. créer de faux comptes, automatiser des activités inauthentiques, générer ou partager du contenu intentionnellement trompeur, manipuler les systèmes de classement, etc.) sont interdites. Il en est de même pour le contournement des restrictions d'accès, d'utilisation ou de la disponibilité des services (ex. en tentant de *jailbreaker*[131] un système d'IA ou en effectuant un scraping non autorisé).

Microsoft fournit quelques exemples d'utilisation abusive de ses services qui ne sont pas autorisés :

- l'obtention ou la tentative d'obtention d'un accès non autorisé à des systèmes sécurisés tels que des comptes, des systèmes informatiques, des réseaux ou tout autre service ou infrastructure ;
- le déploiement ou la tentative de déploiement de logiciels ou de codes de toute nature sur des systèmes non autorisés susceptibles d'affecter négativement le fonctionnement des réseaux, services ou infrastructures ou de ceux d'autres réseaux ou infrastructures ;
- perturber ou tenter de perturber les services de Microsoft ou d'autres entreprises, ou tout autre système, par quelque moyen que ce soit, y compris, mais s'en s'y limiter, les attaques par déni de service ;
- toute tentative visant à contourner ou à entraver l'accès aux services, leur utilisation ou leur disponibilité. Ceci inclut toute tentative de contournement des mesures qui peuvent être prises sur les comptes des utilisateurs.

LinkedIn apporte également des précisions dans les conditions d'utilisation qui lui sont propres. La plateforme LinkedIn n'autorise pas les logiciels tiers, y compris les robots d'indexation, les bots informatiques, les modules et extensions de navigateur qui effectuent du *web scraping*[132], modifient ou automatisent les activités sur le site de LinkedIn.

Plus précisément, la plateforme interdit l'utilisation de bots ou « *d'autres méthodes automatisées non autorisées* » pour accéder aux services, ajouter ou télécharger des contacts, envoyer ou rediriger des messages, créer, commenter, aimer, partager des publications, ou générer de manière générale un engagement artificiel.

Les faux comptes ne sont pas non plus autorisés, tout comme les outils ou services qui essaient de manipuler les algorithmes de LinkedIn.

Enfin, le fait de perturber le fonctionnement du service ou d'imposer une charge disproportionnée sur la plateforme (ex. spam, attaque par déni de service, virus, etc.) est également interdit.

[130] https://www.microsoft.com/fr-fr/digitalsafety/policies (paragraphe sur la mauvaise utilisation des services de Microsoft ; https://www.microsoft.com/fr/servicesagreement (partie sur les codes de conduite) ; https://fr.linkedin.com/legal/user-agreement? (partie 8.2.13 « *Règles* »).
[131] Contourner des restrictions imposées par un fabricant sur un appareil.
[132] Collecte automatique de données sur des sites web au moyen de programmes ou de scripts informatiques.

4. Contenus indésirables (spams)[133]

Microsoft interdit, pour l'ensemble de ses services, d'envoyer des spams, de pratiquer l'hameçonnage et de tenter de créer ou de diffuser des logiciels malveillants.

Il mentionne une liste d'exemples de pratiques de spam interdites :

- envoyer des messages non sollicités aux utilisateurs ou publier des commentaires à caractère commercial, répétitifs ou trompeurs ;
- utiliser des titres, des vignettes, des descriptions ou des balises pour induire les utilisateurs en erreur et leur faire croire que le contenu porte sur un sujet ou une catégorie différente de celle qu'il traite réellement ;
- l'envoi en masse de courriels, de publications, de demandes de contacts, de SMS, de messages instantanés ou de communications électroniques similaires non sollicités ou indésirables ;
- l'utilisation de tactiques trompeuses ou abusives pour tenter de tromper ou de manipuler le classement ou d'autres systèmes algorithmiques, notamment le spam de liens, le *cloaking*[134] ou le bourrage par mots-clés.

Pour son service en particulier, LinkedIn interdit tout comportement qui pourrait perturber son bon fonctionnement ou lui imposer une charge disproportionnée. Cela comprend notamment les spams ou les techniques d'hameçonnage (cf. *supra* sur les comportements inauthentiques et les engagements artificiels).

[133] https://www.microsoft.com/fr-fr/digitalsafety/policies (paragraphe sur les contenus indésirables pour les services de Microsoft) ; https://www.microsoft.com/fr/servicesagreement (partie sur les codes de conduite) ; https://fr.linkedin.com/legal/user-agreement? (partie 8.2.16 « *Règles* »).
[134] Technique qui vise à manipuler le classement des résultats de recherche ou de cacher du contenu malveillant ou indésirable.

1. Les médias synthétiques et manipulés[135]

Pour les contenus organiques, Snapchat mentionne simplement qu'il est interdit de générer des *deepfakes*.

Pour les contenus publicitaires, la plateforme précise qu'elle interdit les contenus qui comprennent des avatars de synthèse ou des ressemblances visuelles ou vocales avec une personne réelle et qui ont été manipulés à des fins frauduleuses ou fallacieuses (que ce soit par le biais d'IA générative ou d'un montage trompeur).

2. Fraude et usurpation d'identité[136]

Snapchat précise que les fraudes et les pratiques trompeuses interdites incluent les contenus faisant la promotion d'escroqueries de toute nature, les programmes d'enrichissement rapide, les contenus payants ou sponsorisés non autorisés ou non présentés comme tels, le marketing à paliers multiples ou les systèmes pyramidaux et la promotion de biens ou de services frauduleux, y compris les produits ou documents contrefaits.

Snapchat interdit également l'usurpation d'identité sur la plateforme.

3. Comportements inauthentiques et engagements artificiels[137]

Dans l'ensemble, Snapchat interdit l'utilisation du service d'une manière qui pourrait déranger, perturber ou affecter négativement les utilisateurs, ou qui aurait comme conséquence d'endommager, de surcharger ou d'altérer le fonctionnement de la plateforme. Par exemple, Snapchat interdit d'utiliser des robots ou tout autre moyen informatisé, ainsi que le recours à des applications tierces, pour accéder à la plateforme.

Le fait de télécharger des virus ou d'autres codes malveillants, de compromettre, contourner ou éviter la sécurité de Snapchat est également prohibé. Dans le même ordre, le fait de tester la vulnérabilité de la plateforme est interdite.

4. Contenus indésirables (spams)[138]

Snapchat interdit :

- le spam, y compris le contenu ou l'engagement non sollicité ou artificiellement gonflé ;
- tout système qui permet d'acheter un accroissement d'abonnés ;
- la promotion d'applications de spam ;
- les publications ou les partages massifs, répétitifs ou fréquents.

---

135 https://www.snap.com/ad-policies? (concernant les contenus publicitaires) ; https://help.snapchat.com/hc/fr-fr/articles/25494876770580-IA-g%C3%A9n%C3%A9rative-sur-Snapchat (pour les contenus organiques).

136 https://values.snap.com/policy/policy-community-guidelines/harmful-false-deceptive-information (voir l'onglet « *vue d'ensemble* » ainsi que le point 2. Fraude) ; https://values.snap.com/policy/prohibited-content/deceptive-content (fraudes et arnaques).

137 https://www.snap.com/terms-02-26-2024 (point 7. Respect des services et des droits sur Snap).

138 https://values.snap.com/policy/policy-community-guidelines/harmful-false-deceptive-information (voir l'onglet « *vue d'ensemble* » ainsi que le point 3. Spam).

1. **Les médias synthétiques et manipulés**[139]

Google Search n'autorise ni contenu audio, vidéo ou image manipulé dans le but de tromper, frauder ou d'induire en erreur en présentant une version mensongère d'actions ou d'événements qui ne se sont manifestement pas produits. Ceci inclut tout contenu susceptible d'induire une personne raisonnable en erreur quant à sa compréhension ou son interprétation, et pouvant ainsi causer un préjudice important à des groupes ou des individus, ou compromettre gravement la participation ou la confiance dans les processus civiques ou électoraux. Google Search a d'ailleurs une page dédiée aux mesures mises en place contre les *deepfakes*[140].

2. **Fraude et usurpation d'identité**[141]

Google interdit le fait, pour un utilisateur, d'usurper l'identité d'une autre personne ou une organisation, ou de se livrer à des activités visant à tromper, frauder ou induire en erreur. Cela inclut, par exemple, le fait de prétendre faussement être affilié à une autre personne ou organisation, ou d'être soutenue par celle-ci.

3. **Comportements inauthentiques et engagements artificiels**

Google Search n'a pas de conditions d'utilisation spécifiques similaires à celles des très grandes plateformes en ligne, qui listent des interdictions générales de comportements inauthentiques ou d'engagements artificiels.

Néanmoins, quelques informations peuvent être trouvées dans les politiques dédiées au spam de Google Search (cf. *infra*, sur les contenus indésirables).

4. **Contenus indésirables (spams)**[142]

La politique de Google Search en matière de lutte contre les contenus indésirables comprend les interdictions suivantes :

- les liens toxiques, c'est-à-dire la création de liens vers ou depuis un site principalement dans le but de manipuler les classements dans les résultats de recherche ;
- l'envoi de requêtes automatiques à Google (trafic automatisé) ;
- les pratiques et logiciels malveillants ;
- les fonctionnalités trompeuses, qui englobent les pratiques consistant à créer intentionnellement des sites qui font croire aux utilisateurs qu'ils peuvent accéder à certains contenus ou services, alors que ce n'est pas le cas (ex. un site qui prétend fournir certaines fonctionnalités, comme la fusion de PDF, mais qui redirige intentionnellement les utilisateurs vers des annonces mensongères plutôt que de fournir les services supposés) ;
- l'utilisation abusive de contenus à grande échelle, c'est-à-dire :

[139] https://blog.google/intl/en-in/company-news/technology/our-approach-to-protecting-users-from-the-risks-of-ai-generated-media/
[140] https://blog.google/products/search/google-search-explicit-deep-fake-content-update/
[141] https://policies.google.com/terms?hl=en-US#toc-using
[142] https://developers.google.com/search/docs/essentials/spam-policies?hl=fr ; https://www.google.com/intl/fr_ALL/search/howsearchworks/how-search-works/detecting-spam/

  - l'utilisation d'outils d'IA générative ou d'autres outils similaires pour générer de nombreuses pages sans ajouter de valeur pour les utilisateurs ;
  - le détournement de flux, des résultats de recherche ou d'autres contenus pour générer de nombreuses pages (y compris par le biais de transformation automatisées telles que les synonymes, la traduction ou d'autres techniques d'obscurcissement), ne fournissant que peu d'intérêt aux utilisateurs ;
  - l'assemblage ou la combinaison de contenus issus de différentes pages web sans ajout de valeur ;
  - la création de plusieurs sites dans le but de masquer la nature des contenus à grande échelle ;
  - la création de nombreuses pages qui contiennent des mots clés de recherche, mais dont le contenu a peu ou pas de sens pour le lecteur ;
- l'utilisation abusive de la réputation d'un site, qui consiste à publier du contenu tiers sur un site hôte, principalement en raison des signaux de classement déjà établis par ce site, qui sont surtout issus de son contenu propriétaire. Elle vise à faire en sorte que le contenu soit mieux classé que s'il n'était pas associé à d'autres contenus ;
- la redirection trompeuse vers une URL autre que celle demandée par un utilisateur.

1. Les médias synthétiques et manipulés

Wikipédia n'a pas de CGU sur les médias synthétiques et manipulés formellement rédigées comme les autres grandes plateformes. Cependant, ses règles communautaires interdisent l'insertion de contenus trompeurs, ce qui pourrait implicitement entrer dans cette catégorie. Cela peut se confirmer grâce au projet communautaire *WikiProject AI Cleanup*[143], qui vise à lutter contre les contenus générés par IA qui sont sans source et de mauvaise qualité. En août 2025, Wikipédia a d'ailleurs adopté une politique permettant aux contributeurs de proposer la suppression rapide des articles suspectés d'être générés par IA.

2. Fraude et usurpation d'identité[144]

La fondation Wikimédia mentionne dans ses conditions générales d'utilisation qu'il est interdit de se livrer à la fraude ou à l'usurpation d'identité. Pour cette dernière, il s'agit notamment de prohiber « *toute fausse déclaration concernant [une] affiliation à une personne ou une entité* » ou « *l'utilisation du nom ou du nom d'utilisateur d'une autre personne dans l'intention de tromper* ».

3. Comportements inauthentiques et engagements artificiels[145]

Contrairement à la plupart des très grandes plateformes en ligne, Wikipédia n'a pas de conditions d'utilisation spécifiquement dédiées aux comportements inauthentiques et aux engagements artificiels.

Cependant, les conditions d'utilisation de la fondation Wikimédia et les politiques de la communauté interdisent certains comportements problématiques comme :

- le vandalisme, défini comme la modification d'une page de manière intentionnellement perturbatrice ou malveillante ;
- l'utilisation de plusieurs comptes pour tromper, induire en erreur, perturber, fausser le consensus ou contourner un blocage ou une sanction ;
- les botnets non autorisés ;
- le fait de tester les vulnérabilités des systèmes ou des réseaux techniques de Wikipédia ;
- le fait d'accéder, sans autorisation, aux systèmes informatiques de Wikipédia, de les utiliser et d'en altérer le fonctionnement.

4. Contenus indésirables (spams)[146]

Sur Wikipédia, le terme « *spam* » est employé pour désigner l'ajout irraisonné de liens externes ou internes, qui ne sont pas autorisés.

[143] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_AI_Cleanup
[144] https://foundation.wikimedia.org/w/index.php?title=Policy:Terms_of_Use (Point 4. « *S'abstenir de certaines activités* »).
[145] https://foundation.wikimedia.org/w/index.php?title=Policy:Terms_of_Use (Point 4. « *S'abstenir de certaines activités* ») ; https://commons.wikimedia.org/wiki/Commons:Blocking_policy/fr
[146] https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Spam

Annexe 2 – Actions de l'Arcom en 2025 en matière de lutte contre la manipulation de l'information en ligne

A. Au niveau européen

1. Participation au groupe de travail du Comité pour les services numériques dédié à l'intégrité de l'espace informationnel

Le Comité pour les services numériques a mis en place 8 groupes de travail afin d'accompagner la mise en œuvre du RSN. Parmi eux, un groupe est dédié à l'intégrité de l'espace informationnel et traite des enjeux suivants :

- processus électoraux (préparation, retour d'expérience) ;
- manipulation de l'information et ingérences étrangères ;
- mésinformation et désinformation ;
- autres questions relatives au discours civique.

Dans le cadre de ce groupe de travail, l'Arcom a participé activement à la conception du *DSA Elections toolkit*, une initiative portée par la Commission européenne et les coordinateurs pour les services numériques (CSN) des États membres. Son objectif est d'améliorer et d'optimiser l'engagement des CSN auprès des grandes plateformes, afin d'atténuer les risques liés aux processus électoraux. Ce document se présente comme une « *boite à outils* » comportant des actions visant à guider les CSN sur l'ensemble du cycle électoral. Il couvre différentes thématiques, adaptables selon les compétences juridiques et les moyens des CSN :

- la mise en réseau ;
- la sensibilisation et la communication au public ;
- la surveillance en matière de conformité et d'analyse ;
- la réponse aux incidents.

2. Participation à la *taskforce* du code de conduite contre la désinformation, en qualité de représentante du Media Board

La *taskforce* du code de conduite contre la désinformation instaure un cadre de coopération visant à permettre le suivi des engagements des signataires, notamment des grandes plateformes y ayant souscrit. Il s'agit de l'enceinte d'échange privilégiée en cas d'activation du *Rapid Response System* (RRS) pour les élections nationales et européennes. Il s'agit d'un mécanisme de signalement et de réaction rapide prévu par le code, qui permet aux organisations de la société civile d'avoir un canal de signalement privilégié avec les grandes plateformes signataires, et d'en échanger sous un très court préavis.

3. Transmission des éléments techniques de VIGINUM à la Commission européenne

Dans son rôle de coordinateur pour les services numériques pour la France, l'Arcom assure la communication des éléments techniques qui lui sont transmis par VIGINUM à la Commission européenne. Ces éléments permettent d'alimenter les procédures formelles que cette dernière a ouvert à l'encontre de plusieurs grandes plateformes.

4. Participation au jumelage avec l'autorité de régulation des médias ukrainienne

Réunissant les autorités de régulation des médias italienne, française, allemande et grecque, cette initiative visait à renforcer les capacités institutionnelles du régulateur ukrainien, en encourageant l'adoption des meilleures pratiques européennes en matière de régulation des médias et de circulation des contenus sur les plateformes en ligne.

Financé par l'Union Européenne (1,5 million d'euros sur une durée de 18 mois), ce projet a donné lieu à divers ateliers, sessions de formation et activités de partage d'expérience, notamment en matière de manipulation de l'information sur les grandes plateformes.

B. Au niveau national

1. Intervention du président de l'Arcom lors du forum organisé par VIGINUM

Le président de l'Arcom est intervenu lors du forum organisé par VIGINUM le 28 mars 2025 pour rappeler les missions historiques de l'Arcom ainsi que l'évolution du cadre juridique en matière de lutte contre la manipulation de l'information et de régulation du numérique de façon plus générale.

Dans ce cadre, le président a proposé trois axes de travail pour l'année 2025 :
- la poursuite des efforts de sécurisation du débat public en coopération avec VIGINUM ;
- l'exploitation et la mise en perspective des rapports de transparence publiés par les grandes plateformes, à travers une mobilisation de la recherche, de la société civile, qui disposent des moyens humains et techniques pour réaliser ces analyses ;
- le renforcement de l'économie des médias producteurs d'information.

2. Intervention des services de l'Arcom dans le cadre des sessions en région organisées par l'Institut des Hautes Études de la Défense Nationale (IHEDN)

En 2025, les services de l'Arcom sont intervenus à trois reprises lors de sessions en région organisées par l'IHEDN, sur la thématique de la régulation du numérique et des enjeux liés à la manipulation de l'information.

3. Réalisation du module sur l'IA et l'information en partenariat avec le Conseil national de l'intelligence artificielle et du numérique (CIANum)

L'Arcom et le CIANum se sont associés en 2025 pour concevoir un module pédagogique sur l'IA et l'information, destiné aux citoyens souhaitant comprendre comment l'IA, notamment générative, transforme l'ensemble de la chaine de l'information, ainsi que les risques et les opportunités qu'elle présente. Il délivre certains conseils pour exercer son esprit critique de manière éclairée, dans un environnement où les techniques de manipulation de l'information se multiplient.

4. Préparation des élections municipales des 15 et 22 mars 2026 en coopération avec les acteurs impliqués dans la campagne électorale

Dans le cadre d'un plan d'action visant à s'assurer de la sincérité des scrutins municipaux, l'Arcom a rencontré une série d'acteurs impliqués dans la campagne électorale (chercheurs, organisations de la société civile, autorités, plateformes en ligne, etc.) afin de préparer ces différentes échéances électorales et d'anticiper la menace informationnelle.