CSA
CONSEIL SUPÉRIEUR DE L'AUDIOVISUEL

THÉMA

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

Bilan des mesures mises en œuvres en 2020

Septembre 2021

Sommaire

Introduction

Remarques générales

Onze opérateurs de plateforme en ligne soumis au devoir de coopération prévu par le titre III de la loi du 22 décembre 2018 relative à la lutte contre la manipulation de l'information se sont livrés, pour la seconde année, à l'exercice de déclaration au CSA des moyens mis en œuvre pour lutter contre la diffusion de fausses informations : Dailymotion, Facebook, Google, LinkedIn, Microsoft, Snapchat, Twitter, Unify, Webedia, la Fondation Wikimédia et Verizon Media.

Le présent bilan se concentre sur **l'évolution des mesures** durant l'exercice 2020, par rapport à 2019, en portant une **attention particulière aux dispositions prises pour répondre au contexte exceptionnel** de la crise sanitaire. Certaines plateformes ont rendu compte dans leur déclaration de l'émergence, à la faveur de cette crise, de nouvelles problématiques en matière de manipulation de l'information et des mesures qu'elles ont mises en œuvre en conséquence. Il atteste des efforts importants et croissants fournis par les opérateurs pour œuvrer à la lutte contre la propagation des fausses informations sur leur(s) service(s), témoignant de leur prise de conscience de l'importance de ces enjeux.

Le Conseil invite le lecteur à se référer au bilan publié l'année passée et aux déclarations des plateformes pour avoir une vision exhaustive des moyens déployés.

Le CSA tient en outre à saluer l'esprit de coopération et la disponibilité dont a fait preuve la grande majorité des opérateurs, ainsi que la richesse du dialogue noué avec eux. Il tient à signaler une progression globalement positive de la quantité et de la qualité des informations déclarées par rapport à l'année passée.

Néanmoins, l'examen du détail des déclarations amène à nuancer fortement ce constat global sur trois points.

D'une part, le niveau de précision de réponses est très hétérogène en fonction des opérateurs. Plus encore que l'année passée, Verizon Media se distingue par une déclaration particulièrement étique, tant sur la forme que sur la quantité des éléments communiqués.

D'autre part, le CSA rappelle que s'il donne la possibilité de lui signaler la confidentialité de certains éléments que les opérateurs ne souhaitent pas voir publiés et notamment ceux soumis au secret des affaires, il estime que certains acteurs y recourent trop abondamment,

refusant ainsi de partager avec le public des informations dont le caractère confidentiel semble parfois discutable (Dailymotion notamment).
Enfin, si des efforts notables ont été faits sur ces sujets, le Conseil en appelle à davantage de coopération de la part de certaines plateformes sur la transparence des algorithmes, la lutte contre la manipulation de l'information dans le domaine publicitaire ou encore dans la fourniture d'un certain nombre de données chiffrées qui auraient permis une meilleure compréhension des enjeux et forces en présence.

Cadre et méthode

Le titre III de la loi du 22 décembre 2018 prévoit un devoir de coopération dans la lutte contre la diffusion de fausses informations pour les opérateurs de plateforme en ligne[1] dépassant un seuil de connexion de **5 millions de visiteurs uniques par mois**, par plateforme, calculé sur la base de la dernière année civile[2].

Dans cet objectif, ces opérateurs sont tenus de mettre en œuvre un dispositif de signalement des fausses informations accessible et visible. Ils doivent également prendre des mesures complémentaires pour améliorer la transparence de leurs algorithmes, pour mieux promouvoir les contenus issus d'entreprises et agences de presse et de services de communication audiovisuelle, contre les comptes propageant massivement de fausses informations, pour améliorer l'information des utilisateurs sur les contenus sponsorisés d'information se rattachant à un débat d'intérêt général et pour développer les actions d'éducation aux médias et à l'information (EMI) en direction de ces derniers.

La loi prévoit que ces mesures et les moyens que les opérateurs concernés y consacrent **soient rendus publics et fassent l'objet d'une déclaration annuelle au CSA.** Elle charge celui-ci de publier un bilan périodique de leur application et de leur effectivité, en vue duquel il peut recueillir les informations nécessaires à son élaboration auprès des opérateurs concernés. Elle lui donne également la compétence d'adresser des recommandations aux opérateurs concernés visant à améliorer la lutte contre la diffusion de fausses informations : à ce titre, le Conseil a publié une recommandation le 15 mai 2019.

[1] Tels que définis en ces termes à l'article L 111-7 du Code de la consommation : « *I.- Est qualifiée d'opérateur de plateforme en ligne toute personne physique ou morale proposant, à titre professionnel, de manière rémunérée ou non, un service de communication au public en ligne reposant sur :*
1° Le classement ou le référencement, au moyen d'algorithmes informatiques, de contenus, de biens ou de services proposés ou mis en ligne par des tiers ;
2° Ou la mise en relation de plusieurs parties en vue de la vente d'un bien, de la fourniture d'un service ou de l'échange ou du partage d'un contenu, d'un bien ou d'un service. (…) ».

[2] Seuil fixé par l'article 1 du décret n° 2019-297 du 10 avril 2019 relatif aux obligations d'information des opérateurs de plateforme en ligne assurant la promotion de contenus d'information se rattachant à un débat d'intérêt général.

Dans une logique de coopération et sur la base de cette recommandation, un questionnaire a été adressé aux opérateurs de plateforme en ligne et mis en ligne sur le site du CSA en janvier 2021. Il intégrait, par rapport à la version de l'année passée, des questions complémentaires et des précisions issues de réflexions conduites avec le comité d'experts sur la désinformation en ligne auprès du CSA[3] et portant sur deux domaines : les mesures en faveur de la transparence des algorithmes et la lutte contre la diffusion de fausses informations dans le domaine publicitaire.

Conformément à la recommandation, le délai de remise de la déclaration était fixé au 31 mars 2021. L'absence de respect de ce délai par un opérateur (Unify) pour le service Doctissimo s'explique par des interrogations d'ordre juridique résolues après échange et réflexion avec le Conseil. En revanche, ce même opérateur n'a transmis aucune déclaration pour sa plateforme Auféminin, pourtant soumise aux obligations en 2020.

Les onze déclarations de Dailymotion, Facebook, Google, LinkedIn, Microsoft, Snapchat, Twitter, Unify, Webedia, la Fondation Wikimédia et Verizon Media sont **rendues publiques** sur le site internet du CSA, expurgées des éléments couverts par le secret des affaires ou indiqués comme confidentiels. Leur instruction a amené le CSA à demander des compléments et des éclaircissements à certains opérateurs, dont il est tenu compte dans l'analyse. En outre, le **comité d'experts sur la désinformation en ligne** auprès du CSA s'est réuni le 6 mai 2021 afin de fournir son analyse, de mettre en perspective les informations reçues et de réfléchir à des pistes de recommandations. Sa contribution a permis de nourrir la réflexion du CSA, dont les conclusions du présent bilan relèvent toutefois de la seule responsabilité.

[3] Voir composition sur le site du CSA : https://www.csa.fr/Informer/Espace-presse/Communiques-de-presse/Regulation-des-plateformes-le-CSA-met-en-place-une-equipe-projet-et-s-entoure-d-un-comite-d-experts-sur-la-desinformation-en-ligne.

1. Les plateformes ayant fait l'objet d'une déclaration

1.1. Présentation des plateformes au vu des déclarations reçues

Onze opérateurs ont fait parvenir au CSA une déclaration pour l'exercice 2020 présentant de manière plus ou moins exhaustive les actions menées en terme de lutte contre la manipulation de l'information sur un ou plusieurs de leurs services. Ces derniers répondent aux deux types d'activités visées par la définition légale des plateformes en ligne :

- classement ou référencement de contenus, biens ou services mis en ligne par des tiers : Bing, Google, Yahoo Search ;
- mise en relation de plusieurs parties en vue de la vente d'un bien, de la fourniture d'un service ou de l'échange ou du partage d'un contenu, d'un bien ou d'un service : Dailymotion, Doctissimo, Facebook, Google, Instagram, Jeuxvideo.com, LinkedIn, Microsoft Advertising, Snapchat, Twitter, Wikipédia, YouTube.

Ces plateformes ont un spectre d'activité large. Plusieurs jouent en priorité un rôle de réseau social généraliste (Snapchat, Instagram, Facebook) ou spécialisé (LinkedIn). D'autres permettent le partage de contenus vidéo et audio à destination de tous les utilisateurs (YouTube, Dailymotion) ou offrent un forum en ligne (Jeuxvideo.com, Doctissimo). Certaines exercent des fonctions de moteurs de recherche (Bing, Google, Yahoo Search) ou référencent des contenus d'information mis en ligne par des tiers (Google, Bing, Yahoo). Enfin, l'une d'elles a une vocation encyclopédique (Wikipédia).

Les services de plateformes sont tous disponibles en France. Certains constituent des versions spécifiques pour le territoire français (Facebook, Bing, Snapchat, Yahoo Search). Si Dailymotion, Webedia (Jeuxvideo.com) et Unify (Doctissimo) sont des entreprises françaises, les autres plateformes sont opérées par des sociétés ou associations établies en Irlande (Facebook Ireland Limited, LinkedIn Ireland Unlimited Company, Google Ireland Limited, Microsoft Ireland Operations Limited, Twitter International Company, Verizon Media EMEA Limited (Yahoo)), aux États-Unis (Wikimedia Foundation Inc.) ou au Royaume-Uni (Snap Inc.), dont certaines ont des bureaux en France (Facebook, Microsoft, Twitter, Snapchat, Google).

Ce sont des acteurs de tailles très différentes. Selon les chiffres déclarés, le nombre moyen de leurs visiteurs uniques mensuels en France serait, les suivants : 46 millions pour YouTube, 98 millions pour Wikipédia[4], 44 millions pour Dailymotion, 22 millions pour Snapchat, 6,3 millions pour Doctissimo, 6,2 millions pour Jeuxvideo.com. Microsoft (Bing et Microsoft Advertising) et Twitter n'ont pas communiqué au Conseil le nombre d'utilisateurs

[4] L'opérateur précise que les utilisateurs de Wikipédia en français ne sont pas nécessairement situés en France.

mensuels en France[5] et les chiffres fournis à ce sujet par Facebook[6], LinkedIn, Verizon Media et Google (pour Google Search) l'ont été de façon confidentielle.

Leur modèle économique est basé pour tout ou partie sur les revenus publicitaires, à l'exception de la Fondation Wikimédia (Wikipédia). Seuls quatre opérateurs ont communiqué au Conseil leur chiffre d'affaires pour l'activité de leur service en France en 2020, dont deux à titre confidentiel : Snapchat, Unify pour Doctissimo, LinkedIn et Verizon Media pour Yahoo Search. Twitter et Facebook ne déclarent pas le chiffre d'affaires des services mais celui de leur filiale française en 2020 (respectivement 13 662 654 € et 616 060 000 €). Dailymotion, Google, Webedia et Microsoft n'ont pas déclaré ces chiffres.

Les données communiquées relatives au nombre de visiteurs uniques, au modèle économique des plateformes et à leur chiffre d'affaires permettent au CSA de mieux tenir compte de leurs spécificités dans l'appréciation des mesures mises en œuvre pour lutter contre la diffusion de fausses informations. Le CSA souhaite que les opérateurs les lui fournissent de façon complète à l'avenir.

1.2. Périmètre des services concernés

La première année d'application de la loi du 22 décembre 2018 avait été l'occasion pour le CSA de s'intéresser à la question du périmètre des services tenus de lui adresser une déclaration au titre du devoir de coopération prévu par le titre III de cette loi.

L'examen des déclarations qui lui ont été faites pour cette deuxième année d'application ainsi que des échanges avec les opérateurs justifient de nouvelles précisions en la matière.

1.2.1. S'agissant de l'assujettissement des opérateurs au devoir de coopération

Des échanges avec certains opérateurs ont cette année montré les questions qui pouvaient se poser s'agissant de leur assujettissement aux obligations posées par la loi du 22 décembre 2018, en particulier en ce qui concerne le devoir de coopération.

À ce sujet, le Conseil considère qu'il ressort des articles L. 163-1 et D. 102-1 du code électoral[7] qu'un opérateur de plateforme en ligne est soumis à la loi sur la manipulation de

[5] Twitter précise le nombre d'utilisateurs actifs quotidiens monétisables au 3e trimestre 2019 : 152 millions.
[6] Facebook communique publiquement, dans sa déclaration, la moyenne du nombre d'utilisateurs actifs de Facebook en France au quatrième trimestre 2020 : 40 millions.
[7] « *I.-Le nombre de connexions au-delà duquel les opérateurs de plateforme en ligne sont soumis aux obligations de l'article L. 163-1 est fixé à cinq millions de visiteurs uniques par mois, par plateforme, calculé sur la base de la dernière année civile* ».

l'information pour une plateforme donnée dès l'instant où *sur la dernière année civile écoulée*, l'activité a dépassé un nombre de connexions depuis le territoire français de cinq millions de visiteurs uniques par mois sur cette plateforme.

Par ailleurs, le Conseil rappelle que sa recommandation n° 2019-03 du 15 mai 2019 invite les opérateurs de plateforme en ligne concernés à lui communiquer une déclaration, au plus tard le 31 mars de *l'année suivant l'année d'exercice sur laquelle elle porte*.

Ainsi, en 2021, les opérateurs de plateformes en ligne concernés par le devoir de coopération avec le Conseil étaient ceux qui, en 2019, avaient dépassé le seuil de cinq millions de visiteurs uniques par mois, par plateforme. Ces opérateurs devaient lui adresser une déclaration au plus tard le 31 mars 2021 au titre des mesures prises en 2020 en matière de lutte contre la diffusion de fausses informations susceptibles de troubler l'ordre public ou d'altérer la sincérité d'un des scrutins visé par la loi.

Ces différentes temporalités peuvent expliquer que **certains opérateurs de plateforme en ligne ayant rencontré des succès d'audience notables au cours des derniers mois ne soient pas mentionnés dans le présent rapport** et, à l'inverse, que des plateformes ayant connu des chutes d'audience en 2020 soient concernées par le présent bilan.

1.2.2. S'agissant de la notion d'opérateur de plateforme en ligne

L'exercice d'instruction des déclarations qui lui ont été adressées cette année a fait ressortir un besoin de précision complémentaire de la part du Conseil s'agissant de la notion « d'opérateur de plateforme en ligne » au sens de la loi du 22 décembre 2018.

Le Conseil considère en particulier que les termes de l'article L. 111-7 du code de la consommation, qui définissent l'opérateur de plateforme en ligne comme celui qui propose à titre professionnel un service de communication au public en ligne ***« reposant sur »*** une activité de type plateforme, doivent s'interpréter comme excluant les opérateurs dont cette activité est l'accessoire de contenus éditorialisés, en particulier les dispositifs de commentaire sous des articles de presse.

À l'inverse, un opérateur qui mettrait à disposition de ses utilisateurs un espace dédié de type plateforme, susceptible d'exister en tant qu'une partie dissociable de son offre de service, répond bien à la définition posée par le code de la consommation. Sous réserve que le nombre de connexions à cette seule partie dissociable du service atteigne le seuil de connexions mentionné par le code électoral, l'opérateur est tenu par l'obligation de déclaration annuelle au Conseil au titre du titre III de la loi du 22 décembre 2018.

Le cas particulier des portails d'information proposant des contenus de tiers. L'exemple de Yahoo Portal.

Dans son rapport 2020, le Conseil avait rappelé que la généralité des termes utilisés par le législateur pour définir les opérateurs de plateformes en ligne concernés par la loi du 22 décembre 2018 et l'absence de critères d'exclusion devaient conduire à considérer que, sous réserve des seuils applicables et des spécificités propres à certains services qui ne correspondraient pas à la définition du code de la consommation, un service d'achat de publicités sur des moteurs de recherches, un magasin d'applications, un espace de partage de vidéos ou une place de marché mettent dans la plupart des cas « *en relation (…) plusieurs parties en vue de la vente d'un bien, de la fourniture d'un service ou de l'échange ou du partage d'un contenu, d'un bien ou d'un service* » et qu'il en allait de même ***« des portails d'information lorsque ces derniers reposent sur la mise à disposition de contenus de personnes physiques ou morales tierces ».***

Ces remarques découlaient notamment de la déclaration de la société Verizon Media, alors propriétaire du portail d'information Yahoo Portal. Cette dernière s'interrogeait sur l'inclusion de son service dans le champ de la loi *relative à la manipulation de l'information* au motif qu'il fonctionnait à partir d'une sélection de sources journalistiques impliquant au moins en partie une intervention humaine et *« ne représentant que peu de risques en termes de manipulation de l'information ».*

Le Conseil constate que les indications qu'il a formulées l'an dernier pour confirmer l'inclusion des portails d'information proposant des contenus de tiers n'ont pas été suivies cette année par la société Verizon Media. L'opérateur a en effet estimé que son portail d'information n'était pas soumis au devoir de coopération prévu par la *loi relative à la manipulation de l'information* au motif qu'il *« n'est pas une plateforme ouverte où les utilisateurs/fournisseurs de contenu peuvent seuls sélectionner et afficher leur contenu ».*

Face à un tel refus de coopérer de la part de la société Verizon Media, le Conseil lui rappelle une nouvelle fois les obligations prévues à l'article 11 de la loi du 22 décembre 2018, qui sont tout à fait compatibles avec l'activité d'un portail d'information, à savoir, d'une part, la mise en place d'un dispositif facilement accessible et visible permettant aux utilisateurs de signaler des fausses informations et, d'autre part, des mesures complémentaires dont l'article 11 de la loi du 22 décembre 2018 fournit une liste illustrative.

2. La mise en œuvre de la coopération entre le Conseil et les opérateurs de plateformes

2.1. S'agissant du questionnaire adressé aux opérateurs préalablement à leur déclaration

L'examen des déclarations a suscité des réflexions autour de la manière dont le Conseil est amené à faciliter le devoir de coopération des opérateurs de plateforme en ligne en matière de lutte contre la manipulation de l'information.

Pour accompagner les opérateurs de plateformes en ligne dans leur obligation de lui adresser une déclaration annuelle sur la mise en place des mesures prévues à l'article 11 de la loi du 22 décembre 2018, le Conseil a pris l'initiative, dès la première année d'application de cette loi, de leur adresser un questionnaire commun.

Ce document simplifie le travail des opérateurs concernés pour fournir des déclarations conformes aux exigences du Conseil. Il offre également l'occasion à ces derniers de présenter au régulateur et au public leurs différentes actions sur des thématiques considérées particulièrement importantes par le Conseil et les experts du sujet de la manipulation de l'information avec lesquels il collabore.

Certains acteurs ont pu regretter que ce questionnaire ne soit pas adapté à leur modèle spécifique et qu'il ne prenne pas en considération la variété des types de plateformes (notamment Snapchat et la Fondation Wikimédia).

Le Conseil tient à rappeler qu'il **veille à prendre en compte la pluralité des modèles des plateformes et l'adéquation des moyens mis en œuvre sur chacune d'entre elles à l'ampleur et à l'impact du phénomène de manipulation de l'information**. Cet enjeu avait été identifié dès sa recommandation du 15 mai 2019 et continue à orienter l'examen des déclarations des opérateurs de plateforme en ligne à toutes les étapes de leur instruction.

Pour autant, afin de pouvoir répondre au mieux à la mission qui lui a été confiée par le législateur de publier un bilan périodique de l'application des mesures prises par les opérateurs de plateformes en ligne et de l'efficacité de ces mesures, le Conseil juge indispensable de pouvoir disposer d'outils capables de lui offrir une vue générale et, le cas échéant, chronologique de ces mesures, ce qu'offre un questionnaire commun à la fois standardisé et évolutif.

Le Conseil considère qu'il revient à chaque opérateur de répondre aux questions qui lui sont applicables dans ce questionnaire. Il n'exclut ainsi pas que des questions se

trouvent sans objet pour certaines plateformes, auquel cas il invite l'opérateur concerné à l'indiquer en apportant les éléments lui permettant de comprendre ce choix**.** Il encourage également les plateformes à identifier et décrire tous les moyens de lutte contre la diffusion de fausses informations qui ne seraient pas abordés dans le questionnaire, lequel ne saurait être exhaustif.

De manière générale, le Conseil appelle chaque opérateur à envisager l'exercice déclaratif que lui impose la loi du 22 décembre 2018 comme un moyen de démontrer au régulateur et au public le caractère éventuellement vertueux de son modèle en apportant, pour ce faire, autant d'éléments circonstanciés possibles en complément et non en substitution des réponses au questionnaire.

2.2. S'agissant de la demande de certains opérateurs de plateformes en ligne de précisions sur le champ d'application des mesures qu'ils sont tenus de prendre

L'examen des déclarations a permis au Conseil d'identifier un besoin des opérateurs de voir précisées certaines définitions contenues dans le cadre réglementaire qui leur est applicable en matière de fausses informations, dont le *« contenu d'information se rattachant à un débat d'intérêt général ».* Sur cette définition précise, le Conseil renvoie aux précisions apportées dans sa recommandation du 15 mai 2019.

Pour le reste, le Conseil constate que les opérateurs de plateformes en ligne concernés par le devoir de coopération sont **tenus à des obligations de moyens qui impliquent des actions vis-à-vis de leur utilisateurs et une transparence à leur égard.** Ce type d'obligations implique plus généralement un ensemble d'acteurs, qui inclut par exemple tout aussi bien des *fact-checkers* que des professionnels de la publicité. Un tel contexte justifie que chacun s'accorde sur le sens des notions clés qu'il emploie ou auxquelles il est confronté, en particulier les utilisateurs des plateformes, dont la navigation peut porter sur plusieurs services.

Aussi, afin de favoriser le travail général de lutte contre la diffusion de fausses informations et l'information éclairée des internautes, le Conseil estime particulièrement important que les opérateurs de plateformes en ligne travaillent à la **constitution de définitions communes et aisément compréhensibles des notions contenues dans leurs politiques de modération.**

2.3. S'agissant de la nécessité de transparence dans la lutte contre la manipulation de l'information

Aux termes de la loi relative à la manipulation de l'information, les opérateurs de plateforme en ligne sont tenus d'apporter au Conseil des informations précisant les modalités de mise en œuvre des mesures prises en application de l'article 11 de cette même loi, et de rendre publics ces mesures et les moyens qu'ils y consacrent. La loi, de même que l'article 58 de la loi du 30 septembre 1986, prévoient en outre que le bilan périodique réalisé par le Conseil soit lui aussi publié.

L'objectif de transparence assigné par le législateur aux opérateurs de plateformes en ligne dépasse donc les rapports que ces derniers entretiennent avec le régulateur. Il vise le public de manière générale.

Dès lors, le Conseil appréhende cette transparence selon trois modalités :

- **l'information fournie par l'opérateur à l'ensemble des utilisateurs de son service** sur les moyens mis en œuvre pour lutter contre la manipulation de l'information. Elle doit être claire, concise et intelligible, et aisément accessible (en étant par exemple délivrée de manière contextuelle et/ou proactive lorsque cela s'y prête) ;
- **la mise à disposition des citoyens et de l'ensemble de la société civile** (journalistes, associations, agents publics, monde associatif, écosystème académique, etc.), de façon plus détaillée, **de toutes les informations publiables** permettant l'analyse de la responsabilité et de l'impact des plateformes dans la dynamique informationnelle ;
- l'obligation de fournir des informations complètes au régulateur, en lui **transmettant l'ensemble des éléments lui permettant la meilleure compréhension possible des mesures déployées**. Il est ici rappelé que la loi[8] donne compétence au Conseil, pour l'accomplissement des missions qui lui sont confiées, de recueillir **toutes les informations nécessaires pour s'assurer du respect des obligations qui sont imposées aux opérateurs**, sans que puisse lui être opposées d'autres limitations que celles qui résultent du libre exercice de l'activité des partis et groupements politiques.

Dans cette même perspective d'information du public, le CSA inscrit sa propre action en matière de lutte contre la manipulation de l'information dans une démarche de transparence renforcée. Soucieux notamment de permettre aux utilisateurs de comprendre les mécanismes mis en œuvre sur des services qu'ils utilisent parfois au quotidien, il veille à

[8] Article 19 de la loi du 30 septembre 1986 relative à la liberté de communication, applicable aux opérateurs de plateforme en ligne conformément à l'article 58 de la même loi.

ce que ces derniers disposent des éléments les plus complets pour être responsables et acteurs de la lutte contre la manipulation de l'information. À cette fin, le Conseil rend public non seulement le bilan des mesures mises en œuvre par les opérateurs mais également les déclarations des opérateurs produites pour l'élaboration de ce bilan[9].

Afin d'améliorer la transparence à l'égard du public, le CSA formule les préconisations suivantes :

- **Fournir de manière proactive aux utilisateurs, sur la plateforme, si possible de manière personnalisée et contextuelle, des explications claires et accessibles** sur les mesures mises en œuvre face aux risques liés à la manipulation de l'information.

- **Faire preuve de plus de transparence vis-à-vis du public en fournissant davantage de précisions chiffrées et d'éléments contextualisés, notamment dans les déclarations**, et communiquer au Conseil toutes les informations, fussent-elles confidentielles, permettant de mieux comprendre les mesures prises et leur impact.

[9] Cela dans le respect des informations légalement protégées que chaque opérateur peut lui signaler comme « confidentielles » lors de sa déclaration, en fournissant les justifications nécessaires, et sans dévoiler, à la demande de l'opérateur, les informations qui permettraient un contournement des outils de lutte contre les phénomènes de désinformation.

3. Les politiques des plateformes en matière de lutte contre la diffusion de fausses informations

Le questionnaire relatif à l'exercice 2020 interrogeait les opérateurs sur leur approche des fausses informations et les définitions qui en découlaient, l'évolution de leur politique de modération en la matière et la manière d'articuler cette dernière avec les exigences de respect de la liberté d'expression.

Au vu à la fois du contenu des déclarations et des règles de communauté inscrites dans les conditions générales d'utilisation (CGU) des différents services, la portée de la notion de fausse information varie largement selon les opérateurs, et leurs approches en conséquence également. Le Conseil en appelle par ailleurs les opérateurs à communiquer davantage de données chiffrées communiquées concernant le nombre de contenus modérés/supprimés, données globalement manquantes dans les déclarations.

DONNÉES CHIFFRÉES SUR L'EXERCICE 2020[10]	
Dailymotion	473 signalements pour fausse information, 10 contenus identifiés comme tel (dont aucune communication commerciale).
Facebook	Suppression de 12 millions de contenus depuis mars 2020 sur Facebook et Instagram liés à la crise sanitaire ; Suppression d'environ 5,8 milliards de faux comptes en 2020.
Google	YouTube : suppression de 9,3 millions de vidéos liées à la crise sanitaire au 1er trimestre 2020 dont (chiffre confidentiel) provenant d'une adresse IP française; suppression de 2 millions de chaînes et 906 millions de commentaires sur l'année 2020. Google : blocage de 9.6 millions d'annonces publicitaires liées au Covid-19 et de 1 500 URL en France.
LinkedIn	24 919 signalements pour fausse information et modération sur 8 703 contenus en France.
Microsoft	Microsoft Advertising : suppression de 1,6 milliards de publicités, retrait de 270 000 sites et suspension de 300 000 comptes.
Snapchat	4 contenus considérés comme étant de « fausses informations » (le reste des chiffres est déclaré de manière confidentielle)
Twitter	168 709 signalements en France pour fausse information susceptible d'altérer un scrutin ou troubler l'ordre public et mesures prises à l'égard de 54 254 comptes dans le monde.
Unify	4 789 signalements pour fausse information, suppression de 1 845 contenus en France (dont aucune communication commerciale) sur Doctissimo.
Verizon Media	Pas de données communiquées.
Webedia	Environ 50 000 fausses informations identifiées en France (évaluation).
Fondation Wikimédia	L'opérateur déclare qu'étant donné son modèle coopératif, Wikipédia n'est pas concerné par ce type de modération.

[10] Dans les cas où ce n'est pas précisé, les chiffres sont déclarés par les opérateurs l'échelle mondiale.

3.1. La « manipulation de l'information », une notion recouvrant plusieurs types de phénomènes

Les plateformes ont essayé d'esquisser, dans leur déclaration, les contours d'une notion encore difficile à appréhender pour le public. Le sens donné à la notion de « *manipulation de l'information* » ou à celle de « *fausse information* » demeure variable d'un opérateur à l'autre, au vu aussi bien de leur déclaration, des règlements de la communauté que des formulaires de signalement.

À titre d'exemple, on peut citer Twitter, LinkedIn ou Snapchat qui, dans leurs règles communautaires, mettent en garde les utilisateurs contre :

- la diffusion de « *pratiques trompeuses et fausses informations susceptibles de causer un préjudice ou malveillantes* » pour Snapchat ;
- la diffusion de « *contenus faux ou trompeurs (…), y compris de fausses informations ou de la désinformation* » rattachés à la période électorale, à la crise sanitaire, au négationnisme ou à la monétisation de ce genre d'informations pour LinkedIn ;
- la manipulation des élections ou l'interférence dans des élections ou dans d'autres processus civiques ainsi que le partage « *des médias synthétiques ou manipulés à des fins de tromperie et susceptibles de causer des préjudices* » pour Twitter.

Facebook indique qu'il s'attache à limiter la diffusion des fausses informations tout en rappelant qu' « *il n'y a qu'un pas entre les fausses informations et la satire ou les avis personnels.* » et définit la désinformation comme « tout contenu faux ou trompeur ».

Cette hétérogénéité tient notamment aux différences d'usages et de menaces rencontrées sur chaque service et aux politiques d'utilisation choisies par les opérateurs, mais également à la complexité d'appréhender la notion de « fausse information ».

Si le législateur a précisé le type de phénomène contre lequel les plateformes doivent lutter (celui de la « *diffusion de fausses informations susceptibles de troubler l'ordre public ou d'altérer la sincérité d'un des scrutins mentionnés au premier alinéa de l'article 33-1-1 de la loi n° 86-1067 du 30 septembre 1986 relative à la liberté de communication »),* les fausses informations en question ne font pas l'objet de plus de précisions[11].

[11] Une tentative par le législateur de définir la notion comme *« toute allégation ou imputation d'un fait inexacte ou trompeuse»*, n'a pas abouti (cf. article 1er du projet de loi adopté en première lecture à l'Assemblée nationale et supprimé en deuxième lecture). Dans son rapport n°677 fait au nom de la commission de la culture, de l'éducation et de la communication, la sénatrice Catherine Morin-Desailly notait *« le cadre ainsi tracé ne permet en aucun cas de distinguer les différentes nuances de « fausse information » des simples erreurs, approximations, ou bien au contraire des informations tout à fait vraies mais dont on ne peut révéler les sources. Il ne paraît en réalité pas possible de ramasser en une définition un ensemble si large, complexe et mouvant ».*

Les termes « *infox* » ou « *fake news* » usités dans le langage courant demeurent généraux et vagues, couvrant aussi bien des informations mensongères que des erreurs ou des inexactitudes. D'autres notions tentent d'appréhender les phénomènes sous l'angle de l'intentionnalité et de la finalité poursuivie : ainsi, le terme de « *mésinformation* » est parfois utilisé pour désigner une **erreur involontaire** dans le relai d'informations. Le vocable de « *désinformation* » (utilisés par certains opérateurs tels que Dailymotion, Google et Facebook) tend quant à lui à désigner l'information induisant **délibérément** en erreur le public par la manipulation de faits ou de récits (sans que ces derniers ne soient d'ailleurs forcément faux).

Ainsi, la difficulté à utiliser un vocabulaire commun pour définir ces pratiques et cibler celles qui doivent faire l'objet de mesures prioritaires de lutte ne favorise pas l'intelligibilité des mesures prises. En ce sens, le **dialogue et la coopération entre les parties prenantes, notamment entre les opérateurs**, ne peut qu'aider à mieux appréhender les phénomènes qui se développent sur les plateformes numériques et les efforts menés pour les juguler.

3.2. Des approches différentes en matière de lutte et de liberté d'expression

De la même manière, les approches en matière de modération et d'éventuelle suppression d'un contenu diffèrent d'un service à un autre. Si Dailymotion, LinkedIn, Snapchat ou Unify (Doctissimo) s'attachent à proscrire toute publication de fausse information définie selon leur propre charte d'utilisation, d'autres approches existent : Twitter a fait le choix de prendre des mesures différenciées selon une typologie de contenus entre « *information trompeuse, affirmation contestée et affirmation non vérifiée* ». À ce titre, Facebook, Google et Microsoft (pour Bing et Microsoft Advertising) indiquent qu'ils ne souhaitaient pas s'ériger en « *arbitres de la vérité* », considérant, d'une part, qu'il n'existe pas de définition universelle des fausses informations et, d'autre part, qu'ils ne sont pas légitimes à décider, sans intervention de tiers, si un contenu diffuse une information certes inexacte mais pouvant relever du débat et de l'expression d'une opinion autour d'une question, ou à l'inverse, s'il est susceptible de porter atteinte à l'ordre public ou d'altérer la sincérité d'un scrutin.

Si l'ensemble des opérateurs prônent la liberté d'expression et le respect des droits fondamentaux au sein de leur déclaration, ils envisagent de manières différentes son articulation avec l'objectif de lutte contre la manipulation de l'information. Ainsi, Dailymotion déclare qu'aucune forme de « bridage » ne peut être envisagée sur sa plateforme, Facebook évoque la notion de « *dommage physique imminent* »[12] ainsi qu'un triptyque « Authenticité, Sécurité, Confidentialité », Twitter souhaite favoriser le contre-discours pour avoir des avis divers sur sa plateforme tout en limitant les contenus pouvant porter atteindre à un scrutin

[12] Notion de « dommage physique » sur laquelle se base d'ailleurs Facebook pour modérer les contenus relatifs à la crise sanitaire.

ou troubler l'ordre public. Google et Microsoft mettent en avant la difficulté d'arbitrer entre la liberté d'expression et l'impératif de lutter contre la diffusion de fausses informations« *même lorsque les internautes les propagent de bonne foi* ».

Dans le contexte de la crise sanitaire qui a marqué l'année 2020 et les mesures qui s'en sont suivies, il apparaît essentiel que les opérateurs s'interrogent sur le risque d'atteinte aux libertés individuelles des utilisateurs par la modération accrue des contenus liés aux enjeux de santé publique[13].

[13] Voir partie 4 sur les mesures prises pendant la crise sanitaire (p.17).

4. Focus : l'impact de la crise de la Covid-19 sur les phénomènes de manipulation de l'information et les réponses apportées

La crise sanitaire qui a marqué l'année 2020 s'est accompagnée de risques majeurs en termes de pratiques de manipulation et de diffusion de fausses informations susceptibles de troubler l'ordre public. Les déclarations des opérateurs font à cet égard mention de la surabondance de ces dernières au sein de leurs espaces d'expression dès l'apparition du virus fin 2019 et au fur et à mesure des épisodes qui ont marqué cette année de pandémie (origine du virus, questions autour des masques, arrivée des vaccins, etc.).

Dès lors, les plateformes ont rapidement mis en place un certain nombre de **dispositifs spécifiques ou complémentaires aux moyens déjà déployés pour lutter contre la manipulation de l'information**, qu'ils présentent dans leur déclaration en réponse aux interrogations formulées par le Conseil dans son questionnaire. Si certaines déclarations de l'année passée présentaient déjà un certain nombre de ces mesures, les éléments recueillis cette année témoignent de l'agilité dont ont dû faire preuve les opérateurs pour s'adapter au fur et à mesure de la crise. Par la force des choses, celle-ci a constitué un véritable **« laboratoire » de l'efficacité des mesures possibles contre la diffusion de fausses informations en situation exceptionnelle.**

Bien que certains acteurs aient indiqué ne pas avoir constaté de profonds changements dans leur modération (la Fondation Wikimédia, Unify, Dailymotion), la crise de la Covid-19 semble avoir eu un réel impact pour les plateformes – notamment pour les réseaux sociaux et les moteurs de recherche - qui ont mis en place de nombreuses mesures et multiplié les collaborations avec les acteurs institutionnels et les professionnels de santé.

4.1. Présentation des mesures sur les services

Mesures générales d'information	Lancement d'un site dédié, accès à la liste des vaccins et lieux de vaccination disponibles dans sa région : Google Lancement d'un site dédié : Facebook Soutien en matière de santé mentale : Snapchat, Facebook
Dispositif de signalement	Modifications des politiques de modération : LinkedIn, Twitter Procédure d'instruction spécifique : Facebook[14], YouTube, Dailymotion Priorisation des signalements : Dailymotion, Snapchat.

[14] Facebook considère les fausses informations liées à la crise sanitaire comme pouvant causer un dommage physique.

Traitement des signalements	Équipes en télétravail avec un impact sur la modération humaine : Facebook, Google, Twitter ; recours accru à la technologie pour la modération : YouTube, Twitter
Politique de modération des contenus	Suppression d'un nombre important de contenus relatifs au Covid-19 : Facebook[15], YouTube[16], Twitter[17] Création d'un libellé apposé sur les messages pour avertir d'une potentielle fausse information en matière de Covid-19: Twitter
Promotion des entreprises et agences de presse et services de communication audiovisuelle	Collaboration avec des entités institutionnelles et de santé de référence : Facebook, Google, LinkedIn, Snapchat, Twitter Mise en place d'un comité d'experts du secteur de la santé dédié à la lutte contre les fausses informations : Doctissimo Mise en place de centres d'informations, bandeaux, panneaux dynamiques etc. dédiés aux informations sur la crise du Covid-19 : Facebook, Google, LinkedIn, Bing, Snapchat, Twitter, Doctissimo, Jeuxvideo.com, Yahoo Search
Collaboration avec des experts	Extension du programme de *fact-checking* : Facebook (avec 1 milliard de subventions à l'International Fact-Checking Network), Google, Bing (nouveau module de *fact-checking Newsguard* proposé aux utilisateurs).
Diffusion massive de fausses informations	Coopération avec les autorités judiciaires, les chercheurs en cybersécurité ou des entités institutionnelles : Facebook, Google Suspension de comptes pour suspicion d'attaques coordonnées : Google
Contenus publicitaires	Labellisation de contenus : Facebook[18], Google Interdiction des publicités n'émanant pas d'autorités de santé, d'ONGs reconnues ou de sources gouvernementales concernant le Covid-19 : Facebook, Google, Microsoft Advertising, Snapchat, Twitter, Yahoo Search Création de campagnes pour promouvoir les informations d'autorité : Facebook, Google Financement des publicités pour les messages d'intérêt public : Google, Facebook Mise à jour de nouvelles règles à l'attention des annonceurs : Google

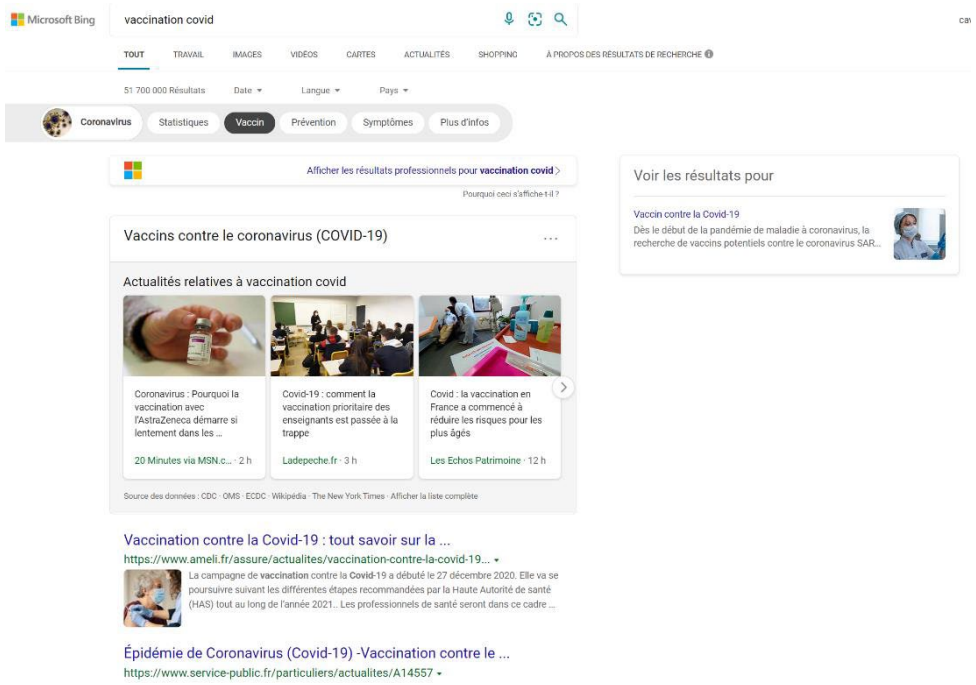
[15] 12 millions de contenus supprimés sur Facebook et Instagram depuis mars 2020 dans le monde.
[16] Vidéos supprimées sur YouTube entre février 2020 et 2021 : déclaré à titre confidentiel.
[17] 2 710 contenus supprimés ; mesures à l'encontre de 4568 comptes dans le monde de janvier à juin 2020.
[18] 167 millions de contenus labellisés par Facebook dans le monde entre mars et octobre 2020, dont 50 millions sur le seul mois d'avril.

Éducation aux médias et collaborations avec le monde de la recherche	Collaboration avec des experts et chercheurs du monde de la santé : Facebook, Google, LinkedIn Campagnes d'EMI spécifiques : Facebook, Google, LinkedIn, Twitter Lettre d'information aux utilisateurs : Dailymotion, LinkedIn, Doctissimo Mise à disposition de données pour étudier les phénomènes de désinformation autour de la crise sanitaire : Twitter

4.2. L'information des utilisateurs au cœur des dispositifs mis en place par les plateformes

La majorité des plateformes, y compris certaines ayant leurs propres équipes rédactionnelles (LinkedIn) ou sélectionnant préalablement les partenaires médias (Snapchat), a œuvré à faciliter la promotion de contenus faisant autorité dans le cadre de la lutte contre l'épidémie. Google, Facebook, Twitter ont mis en **place des fonctionnalités permettant de rediriger les utilisateurs vers les sites officiels ou faisant remonter prioritairement les contenus émanant de sources officielles.** La majorité a créé une ou des pages réunissant des contenus faisant autorité sur le sujet de la crise sanitaire, permettant d'étendre la portée des communications gouvernementales et internationales.

Hub médiatique sur le service Bing[19] :



[19] Source : déclaration de l'opérateur.

Les opérateurs ont également mis en place plusieurs campagnes d'éducation aux médias afin de sensibiliser les utilisateurs aux nombreux dangers liés à la crise. LinkedIn a fait appel à des influenceurs pour partager des contenus d'autorité, Facebook a lancé des vidéos éducatives en coopération avec l'AFP et une campagne publicitaire avec l'OMS, Google a proposé de nouvelles fonctionnalités au niveau local pour connaître les lieux de vaccination et Twitter a démarré un partenariat avec l'UNESCO sur la désinformation en temps de pandémie. Le Conseil remarque ainsi que la crise sanitaire a largement poussé les opérateurs à s'impliquer davantage dans la sensibilisation des utilisateurs aux risques entourant la manipulation de l'information.
Parmi les initiatives notables, figurent également la mise en place d'un comité d'experts de la santé dédié à la lutte contre les fausses informations par Unify sur Doctissimo, la création de filtres par Snapchat pour soutenir les utilisateurs susceptibles de traverser une crise de santé mentale et la mise à disposition des utilisateurs de Bing (Microsoft) d'un outil de *fact-checking* (en collaboration avec *Newsguard*). Le lecteur est invité à se référer à la partie 6 du présent bilan qui rend compte des nombreuses initiatives en la matière.

Bannière mise en place sur le service Jeuxvideo.com[20] **:**

4.3. Un travail de collaboration avec les acteurs centraux de la crise

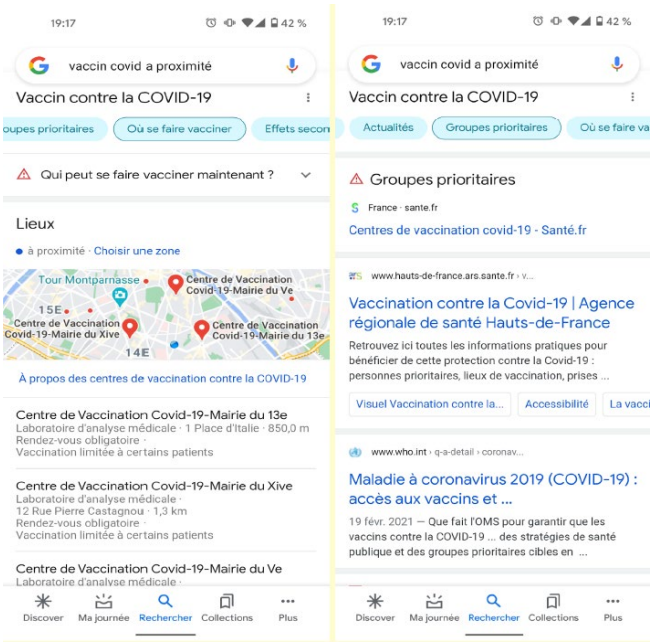
Afin de contrer les fausses informations circulant sur leurs espaces, les plateformes ont également **multiplié les collaborations avec les acteurs gouvernementaux** (Service d'information du Gouvernement, Ministère de la santé) et institutionnels (UNICEF et UNESCO notamment) et les autorités de santé (l'OMS).

Plusieurs plateformes (confidentiel) ont par ailleurs financé les espaces publicitaires pour les campagnes d'informations gouvernementales.

À travers son programme *Google News Initiative*, Google a également financé un hub médiatique afin de regrouper les informations relatives à la vaccination, notamment sur Google Maps, et faciliter le travail de vérification de l'information.

[20] Source : déclaration de l'opérateur.

Fonctionnalités mises en place par Google pour la vaccination[21] :



Enfin, on peut noter la coopération de Facebook et Google avec les autorités judiciaires et des chercheurs en cybersécurité dans le cadre de la lutte contre les attaques coordonnées ainsi que la mise à disposition, par Twitter**,** d'un point d'accès gratuit à ses données dans son API consacré à la COVID-19 afin de permettre à des chercheurs et développeurs d'étudier les pratiques de manipulation de l'information autour de la crise.

4.4. Une difficile articulation entre modération et liberté d'expression ?

La crise aura eu des conséquences considérables sur le nombre de fausses informations au sein des plateformes et a entraîné un travail accru de modération en conséquence sur leurs services (Twitter, Google, Facebook, Doctissimo et Bing notamment).

Google a indiqué avoir restreint plus de 9,6 millions d'annonces publicitaires en France en 2020 liées à des fausses informations sur la crise sanitaire ; Microsoft a empêché plus d'1,7 millions de soumissions d'annonceurs dans le monde et Twitter a restreint plus de 18 000 tweets sponsorisés, ainsi que 4 568 comptes dans le monde de janvier à juin 2020. Pour le service YouTube, entre février 2020 et février 2021, ce sont plus de (chiffre confidentiel) vidéos mises en ligne à partir d'une adresse IP en France et relatives à la désinformation liées à la pandémie de la Covid-19 qui ont été supprimées. Enfin, depuis mars 2020, Facebook a supprimé plus de 12 millions de contenus en rapport avec la crise (sur Facebook et Instagram).

[21] Source : déclaration de l'opérateur.

On peut également ajouter à ces chiffres la montée des attaques coordonnées indiquée par Google qui déclare avoir suspendu 1 500 comptes d'annonceurs en France pour « *tentative de contournement* », notamment des annonces liées à l'urgence sanitaire.

Certains opérateurs ont expliqué dans leur déclaration que la crise avait eu des conséquences directes sur les équipes de modération de leurs services (YouTube et Twitter) et ont déclaré avoir recouru à davantage de modération automatique pour lutter contre la hausse du nombre de fausses informations. Dès lors, se pose la question de la **difficile articulation entre la modération qui est nécessaire en temps de crise pour des enjeux évidents de santé publique et la liberté d'expression qui pourrait être mise à mal par une modération massivement automatisée ne faisant pas intervenir de décision humaine, avec des risques de nombreux faux positifs en l'absence de prise en compte par un humain du contexte dans l'analyse d'un contenu.**.

Ce choix a été rendu nécessaire par le confinement et le télétravail des équipes de modération. Il est assumé chez certains, tel que Google qui déclare qu'une « *option consistait à utiliser nos systèmes automatisés pour appréhender un spectre plus large de contenus afin que la plupart des contenus susceptibles de nuire à la communauté soient rapidement supprimés de YouTube, tout en sachant que de nombreuses vidéos ne feraient pas l'objet d'une revue humaine, et que certaines de ses vidéos qui ne violaient pas notre règlement seraient supprimées. Parce que la responsabilité est notre priorité absolue, nous avons choisi cette seconde option - en utilisant la technologie pour effectuer une partie du travail normalement effectué par les examinateurs. Le résultat a été une augmentation du nombre de vidéos supprimées de YouTube ; plus du double du nombre de vidéos que nous avions supprimées au cours du trimestre précédent.* »

Le Conseil remarque que le caractère exceptionnel de ce type de mesures n'a pas été affirmé dans les autres déclarations, à l'exception de Twitter. Il estime que les celles-ci auraient été plus transparentes et plus complètes si elles avaient fourni des chiffres sur les conséquences de ce recours accru à des systèmes de modération automatisés (taux d'erreurs, nombre de recours, nombre de recours conduisant à une annulation de la décision de modération).

5. Analyses des moyens déployés par les plateformes pour lutter contre la diffusion de fausses informations

5.1. Le dispositif de signalement de fausses informations susceptibles de troubler l'ordre public ou d'altérer la sincérité d'un scrutin

Aux termes de la loi, les opérateurs sont tenus de mettre en place un **dispositif facilement accessible et visible**, permettant à leurs utilisateurs de signaler de **fausses informations susceptibles de troubler l'ordre public ou d'altérer la sincérité du scrutin**, notamment lorsque celles-ci sont issues de contenus promus pour le compte d'un tiers.

Sur ce dispositif, les déclarations diffèrent peu de celles concernant l'exercice 2019. Les mesures s'inscrivant dans la continuité de l'année précédente, certaines plateformes renvoient aux réponses fournies l'année passée (Google, Verizon Media). L'analyse suivante sera concentrée sur les points les plus saillants ainsi que sur les quelques évolutions apportées aux dispositifs préexistants.

Les déclarations font apparaître que cette obligation vient parfois se confronter à la politique générale des plateformes (exposées pour la plupart dans les règles de la communauté). À cet égard, elles sont nombreuses (telles que Google, Facebook, Microsoft) à invoquer la **nécessité de trouver un point d'équilibre entre la modération des contenus, le respect de la liberté d'expression et la libre circulation des informations.**

5.1.1. Une hétérogénéité dans l'accessibilité et la visibilité de l'outil de signalement

À l'instar de ce qui avait déjà été constaté l'année précédente, les plateformes affichent une volonté de développer une expérience utilisateur spécifique, adaptée à la réalité de leur fonctionnement. L'ergonomie des dispositifs reste ainsi variable.

- **Accessibilité et visibilité de l'outil de signalement**

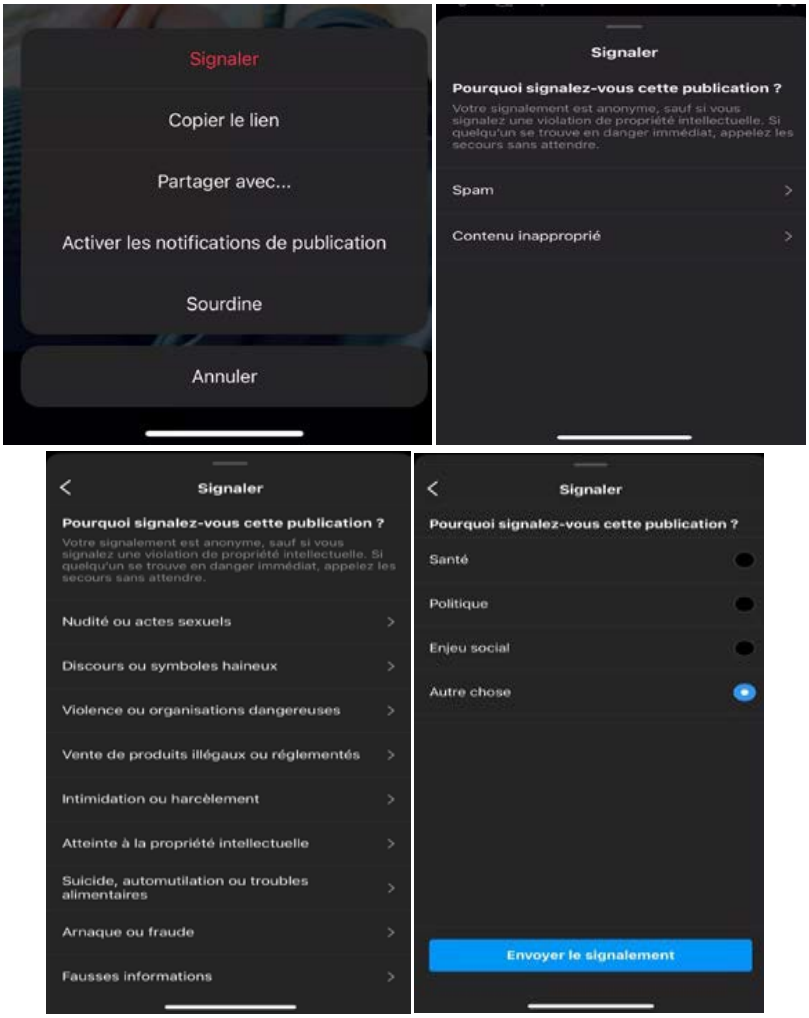
Tous les opérateurs ont mis en place un dispositif de signalement – à l'exception de la Fondation Wikimédia, dans la mesure où le mode de fonctionnement de son service, basé sur l'actualisation constante des contenus par les contributeurs et la modération participative, vise à se substituer à ce type d'outils. Certains optent pour une catégorie spécifique (Facebook, Dailymotion, LinkedIn, Twitter, Snapchat[22], Doctissimo), d'autres rattachent les signalements de fausses informations à la catégorie « *autres* » (Bing,

[22] Dans les espaces *Map*, *Discover* et *Spotlight*.

Jeuxvideo.com, le moteur de recherche Google) tandis que YouTube a prévu de les classer parmi les « spams ou contenus trompeurs ».
Le CSA a encouragé les opérateurs à simplifier l'outil afin que l'envoi d'un signalement puisse être finalisé en **trois clics maximum.**

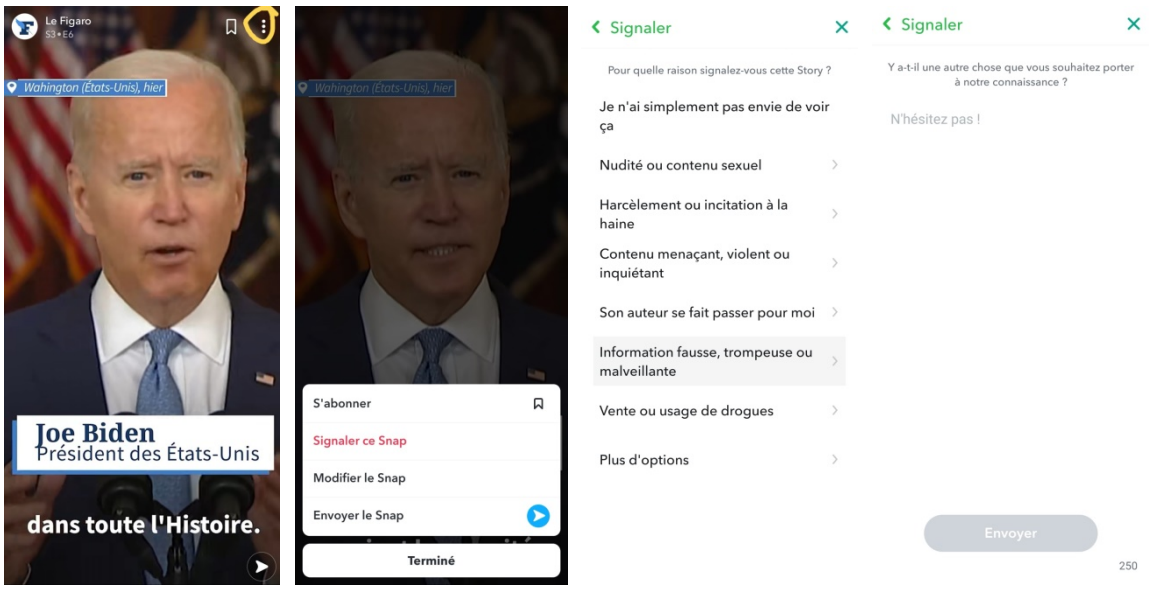
Facebook, LinkedIn ou Dailymotion se sont conformés à cette recommandation, alors que quatre clics sont nécessaires afin de signaler une fausse information sur Snapchat et cinq sur Doctissimo et Twitter. Ce dernier indique par ailleurs avoir tenu compte des recommandations du CSA en mettant en place un mécanisme de signalement directement accessible au niveau du tweet, en un clic et identifié par un drapeau. **Globalement, les outils de signalements prévus par la majeure partie des réseaux sociaux sont plutôt satisfaisants, car situés à proximité du contenu, visibles, facilement accessibles et pour la plupart, aisément compréhensibles.**

Dispositif de signalement Instagram[23] :



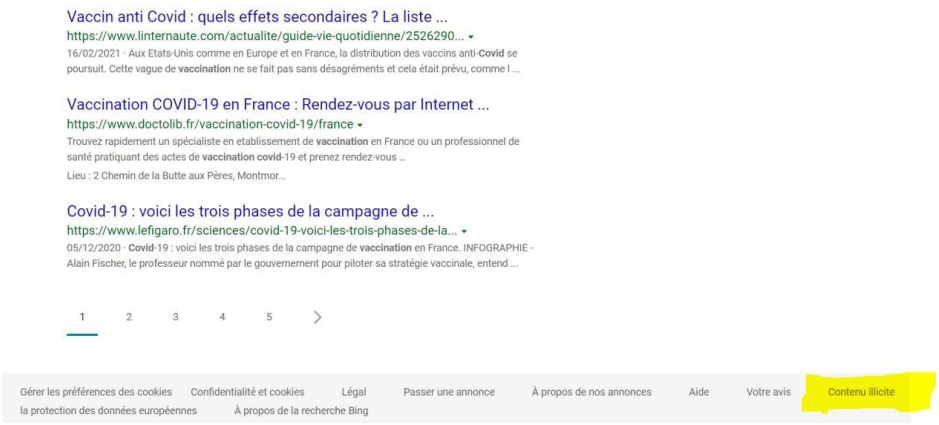
[23] Source : déclaration de l'opérateur.

Dispositif de signalement Snapchat[24] **:**



En revanche, la procédure s'avère complexe pour les moteurs de recherche. Le CSA avait eu l'occasion de constater que les outils de page renvoyant vers un formulaire de demande légale prévus par ces derniers étaient moins intuitifs et moins visibles. Leur chemin d'accès peut s'avérer dissuasif pour l'utilisateur, surtout lorsque ces formulaires, dont la terminologie demeure technique, ne prévoient pas de catégorie « fausse information ». À cet égard, le CSA avait invité les moteurs de recherche à optimiser leurs mécanismes de signalement des fausses informations afin qu'ils soient faciles d'accès et d'utilisation. Or, **aucune amélioration significative** ne semble avoir été apportée par les services concernés au regard des déclarations relatives à l'exercice 2020.

Dispositif de signalement de Bing[25] **:**



[24] Source : observations CSA, août 2021.
[25] Source : déclaration de l'opérateur.

Dispositif de signalement de Google Search[26] :

Il convient de rappeler que l'article 11 de la loi du 22 décembre 2018 rend obligatoire non seulement **l'existence d'un formulaire de signalement des fausses informations, mais également son caractère « *facilement accessible et visible* » pour l'utilisateur.**

- **Pertinence des libellés employés**

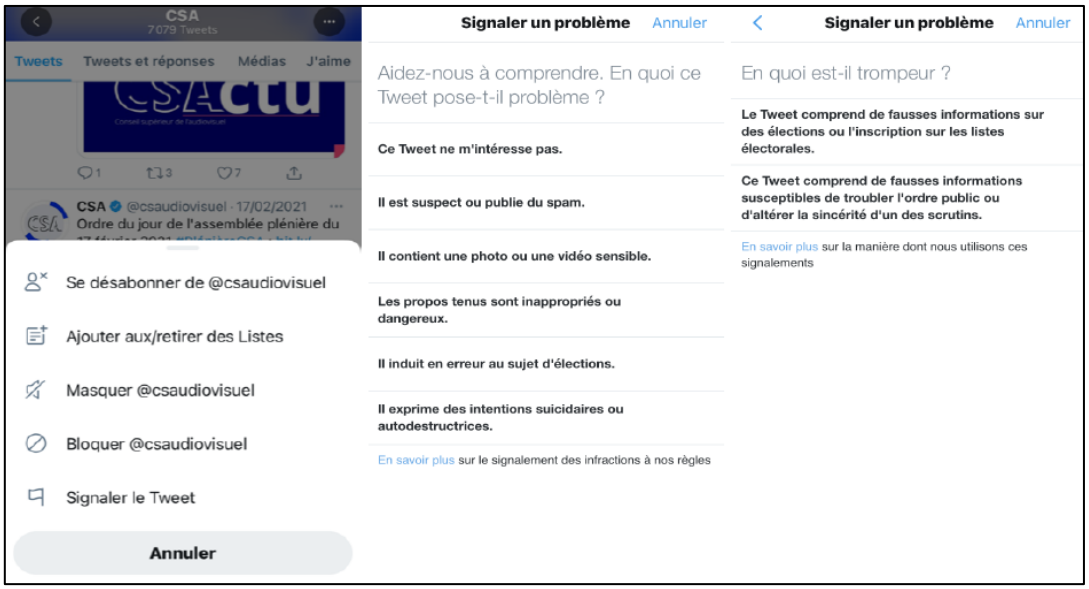
Les termes utilisés pour les motifs de signalement peuvent ne pas toujours apparaître pertinents et peuvent ainsi rendre malaisée leur compréhension par l'utilisateur qui peut être découragé pour aller au bout du signalement. Les services Google, Bing et Jeuxvideo.com échappent, de fait, à cet examen, n'ayant pas prévu de catégorie dédiée aux fausses informations *(cf. supra*).

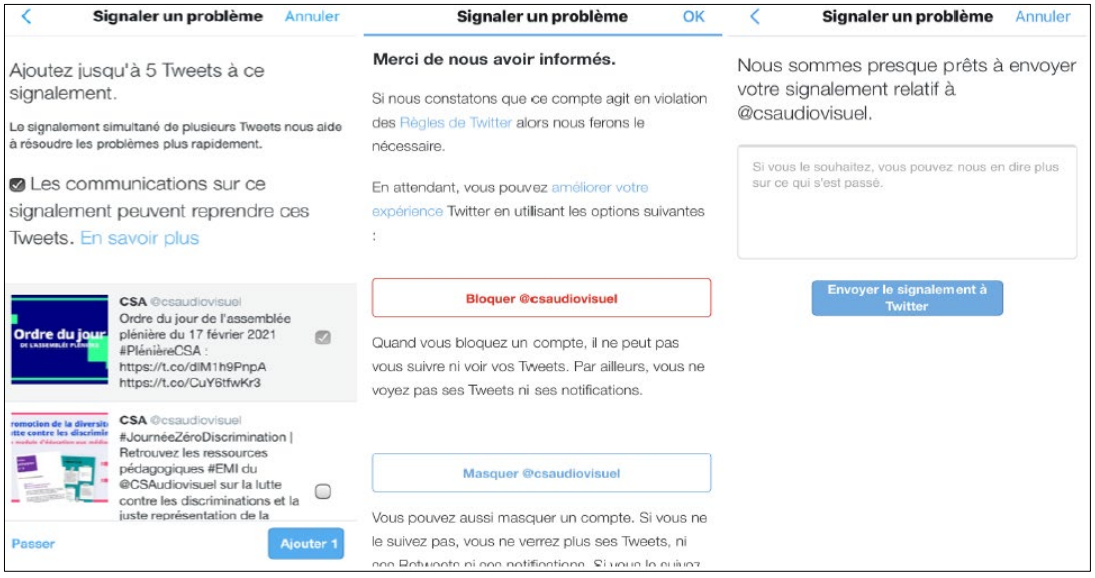
YouTube englobe les fausses informations dans la catégorie « *spam ou contenu trompeur* ». Jeuxvideo.com considère qu'un motif de signalement intitulé « *fausse information* » serait susceptible d'augmenter considérablement le taux de signalements rejetés, car portant sur des contenus n'étant pas considérés comme illégaux ou contraires à leur charte des forums.

Il convient de noter que Twitter a mis en place, en 2020, un outil permettant de signaler les fausses informations susceptibles de troubler l'ordre public ou d'altérer la sincérité du scrutin. Les libellés sont clairs : « *Ce tweet comprend de fausses informations sur des élections ou un autre événement civique* » et « *Ce tweet comprend de fausses informations susceptibles de troubler l'ordre public ou d'altérer la sincérité d'un des scrutins* ». Ils sont disponibles en trois clics et peuvent être complétés par des commentaires et l'ajout d'autres tweets.

[26] Source : observations CSA, août 2021.

Dispositif de signalement Twitter[27] :





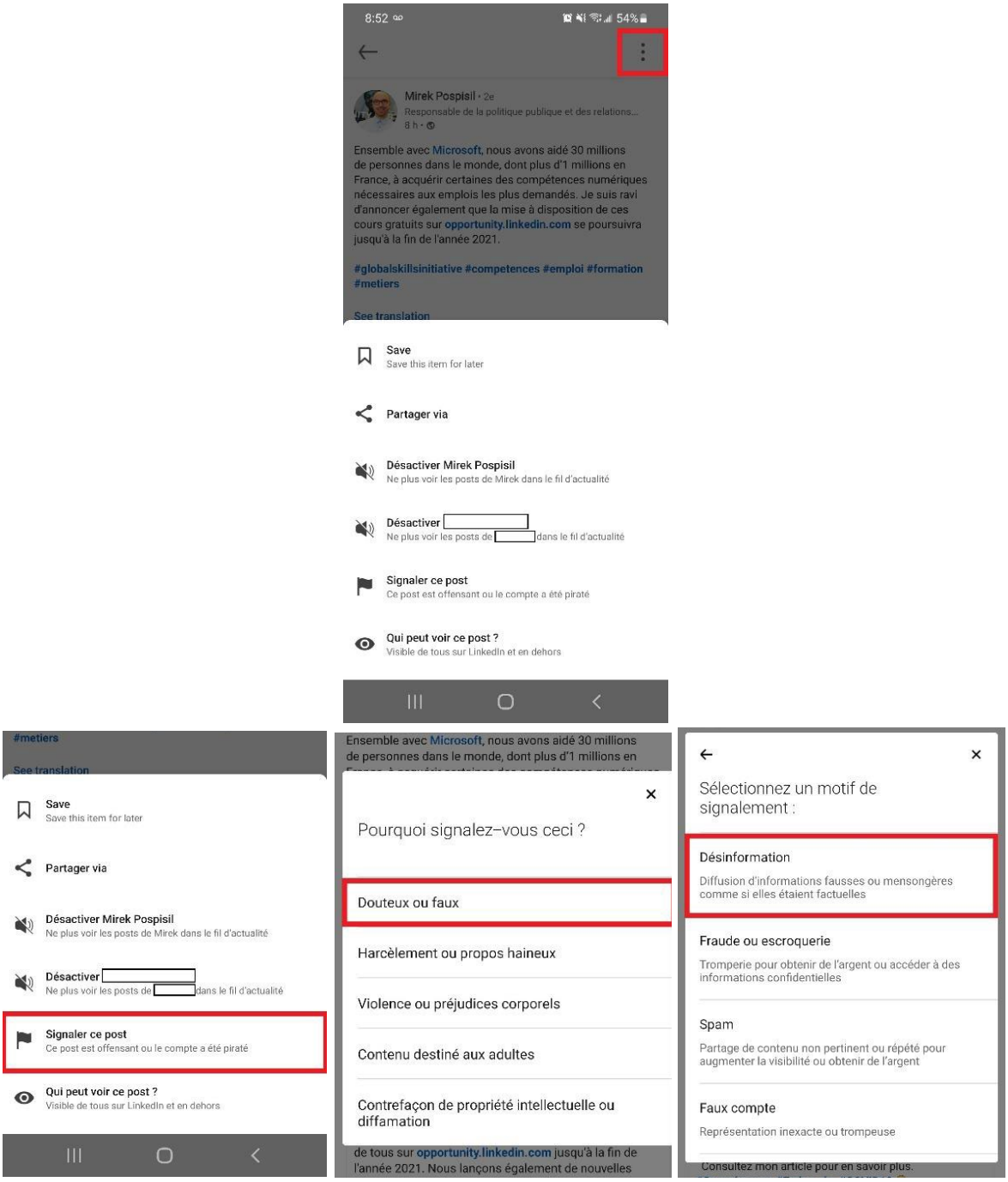
Si la majorité des réseaux sociaux a pris en compte les recommandations du CSA en prévoyant un libellé « fausse information », « désinformation » ou équivalent, le Conseil remarque que certains ne le rendent visible qu'au bout de plusieurs clics, au terme d'un cheminement complexe et non-intuitif. Sur Snapchat, en 2020, il fallait choisir le motif « *Plus d'options* » avant de pouvoir cliquer sur « *Signaler de fausses informations* »[28].

Sur LinkedIn, la possibilité de cocher la case « *Désinformation* » vient après celle de choisir le motif « *Douteux ou faux* ». L'usager est alors mis face à une diversité de motifs, celui des

[27] Source : déclaration de l'opérateur.
[28] Ce n'est plus le cas désormais, le motif figurant dès la première liste des motifs de signalement.

fausses informations pouvant se retrouver dans une catégorie dont la portée est difficile à appréhender**. La place réservée aux fausses informations dans l'arborescence du dispositif pourrait ainsi être regardée comme un indice de l'importance accordée, par la plateforme, à la lutte contre les phénomènes qui en découlent**

Dispositif de signalement LinkedIn[29] **:**



[29] Source : déclaration de l'opérateur.

Par conséquent, on ne peut que considérer avec prudence la déclaration d'opérateurs qui feraient état d'une absence totale de signalement pour fausse information relevant de la loi du 22 décembre 2018 en 2020 (comme c'est le cas de Google pour le présent exercice), eu égard à la pertinence des termes utilisés dans le dispositif de signalement ainsi qu'à la définition que l'opérateur donne de cette catégorie.

5.1.2. Le système hybride privilégié par les plateformes pour la modération des contenus

L'ensemble des plateformes (sauf Unify, Webedia et la Fondation Wikimédia) a recours à la **combinaison d'outils automatique et d'un examen humain pour la détection et le traitement des fausses informations.** Twitter a déclaré axer sa stratégie de modération autour de la technologie pour détecter et supprimer les fausses informations, tout en la couplant avec des moyens humains en forte augmentation. L'opérateur l'explique notamment par la multiplication de ces dernières liée à la crise sanitaire et la nécessaire identification des tweets trompeurs de manière proactive plutôt que de s'en remettre uniquement aux signalements des utilisateurs.

La majorité d'entre elles indiquent que la supervision humaine est essentielle, en particulier pour une modération fine, la décision de retrait des contenus, la suspension/fermeture d'un compte ou la réception et la gestion des voies de recours des auteurs de contenus modérés ou retirés.

Le Conseil peut comprendre le recours accru à la modération automatique en période de crise aiguë, face à l'afflux d'informations sujettes à caution et à l'impact de la pandémie sur l'organisation du travail des équipes de modération. Il estime toutefois indispensable que dans le domaine de la manipulation de l'information, la modération des contenus continue de faire intervenir une décision humaine dès lors qu'elle est susceptible de conduire à la suppression ou à la réduction de la visibilité d'un contenu dont la qualification peut s'avérer délicate, demandant un examen fin du contexte.

En effet, **la modération automatique seule peut conduire à de graves atteintes à la liberté d'expression des utilisateurs** du fait de potentiels faux positifs, *a fortiori* dans des domaines où le contexte est essentiel pour qualifier le contenu. Le Conseil en appelle à davantage à d'informations chiffrées sur ces derniers[30] afin d'évaluer la pertinence d'un recours majoritaire à des algorithmes de recommandation prévalant sur la modération humaine.

[30] Voir partie 5.2 sur la transparence des algorithmes (p.31).

Au vu de ces éléments d'analyse, le CSA formule les préconisations suivantes :

- Pour les moteurs de recherche (Google Search, Yahoo Search et Bing), **améliorer la visibilité et la facilité d'utilisation de leur dispositif de signalement.**

- **Mieux informer** les auteurs de signalement, et utilisateurs ayant publié un contenu signalé, de **l'avancée des procédures de traitement des signalements en cours, et leur en communiquer l'issue dans un délai raisonnable.** Par ailleurs, il apparaît impératif de **mieux expliquer aux utilisateurs les voies de recours existantes.**

- À l'exception de la lutte contre certaines pratiques aisément détectables automatiquement, **maintenir une intervention humaine dans le processus de décision d'une action à l'égard d'un contenu ou d'un compte**.

5.2. Transparence des algorithmes

En matière de manipulation de l'information, deux types de systèmes algorithmiques entrent tout particulièrement en jeu : les systèmes de recommandation de contenus et les systèmes utilisés à des fins de modération.

Les premiers ont un rôle majeur sur la plupart des plateformes dans ce que voit l'utilisateur: ils ont en effet pour fonction de sélectionner, ordonnancer et prioriser, de façon individualisée ou non, le contenu qui proposé à ce dernier. Tout ou partie de ces étapes peuvent être automatisées. Il n'est ainsi pas exclu que des contenus de désinformation soient sélectionnés voire mis en avant, que le système ait été conçu ainsi ou qu'il s'agisse d'effets indésirables (biais).

Les seconds visent à détecter, prioriser voire traiter les contenus préjudiciables ou illégaux. Dans ces systèmes également, tout ou partie de la modération de ces contenus peut être confiée à des systèmes automatiques. Ils viennent ainsi se substituer ou compléter le travail des modérateurs humains.

Les systèmes de recommandation et de modération reposant sur un ou plusieurs algorithmes utilisés par les opérateurs de plateformes en ligne peuvent être liés, notamment lorsque l'opérateur réduit la visibilité d'un contenu signalé ou modéré à l'aide d'un système de recommandation. Ils ont en commun plusieurs caractéristiques : afin d'effectuer la tâche ou série de tâches pour laquelle ils ont été programmés, ils utilisent des données en entrée et produisent un résultat en sortie, telle que la recommandation du prochain contenu à visionner ou la détection d'images similaires à celles qui ont déjà été modérées dans le passé. Les données en entrée ne contribuent pas toutes de manière égale au résultat. Leur

pondération est ajustée selon l'importance que les opérateurs leur donnent individuellement au regard du résultat qu'ils recherchent.

5.2.1. Des systèmes algorithmiques de modération dédiés à la lutte contre la manipulation de l'information qui se multiplient mais dont les effets concrets restent peu documentés

En 2020, les opérateurs ont continué de développer des systèmes algorithmiques de modération spécifiques à la lutte contre la manipulation de l'information ou y contribuant, qu'ils s'attachent à décrire dans leur déclaration. Les systèmes qui ont été déclarés peuvent être regroupés en quatre catégories :

- **L'utilisation de systèmes algorithmiques dans la détection et pour le traitement des signalements de fausses informations** (Dailymotion, Snapchat, Twitter) : les signalements sont remontés algorithmiquement par les systèmes de Dailymotion, qui les traite de manière prioritaire au même titre que ceux liés à du terrorisme ou à de la pédopornographie, *« en particulier durant les périodes électorales et d'urgence »* afin de *« réduire drastiquement leur délai d'instruction »*. Snapchat a ajouté des termes liés à la pandémie mondiale dans son outil de détection des termes abusifs et Twitter utilise ses outils pour identifier les contenus avant modération humaine ;

- **La lutte contre les campagnes coordonnées de manipulation de l'information** (Facebook, Twitter et la Fondation Wikimédia[31]) : la détection de ces comportements par Facebook repose en partie sur des systèmes automatiques. Twitter fait de même et cible notamment les *« manipulations de la plateforme »* qui relèvent d' *« une activité coordonnée, qui tente d'influencer artificiellement les conversations par l'utilisation de comptes multiples, de faux comptes, de l'automatisation et/ou de scripts »* et d' *« une activité coordonnée nuisible qui encourage ou promeut un comportement qui viole les Règles de Twitter »*. La Fondation Wikimédia déclare travailler sur un algorithme de détection de « *faux-nez* » abusifs, une pratique consistant à multiplier les faux comptes sur son service *;*

- **La détection d'hypertrucages ou *deepfakes*** (Facebook, Microsoft) : Facebook utilise des réseaux adversaires génératifs (GANs) pour créer des hypertrucages, permettant ensuite d'entraîner des algorithmes d'apprentissage profond à mieux les détecter (EfficientNet). Microsoft a indiqué donner accès à des outils de détection de *deepfakes* comme Microsoft Video Authenticator et dont certains sont intégrés à son service de *cloud computing*, Microsoft Azure. L'opérateur travaille également au développement

[31] Voir ci-après, partie 5.4 (page 46).

de l'authentification des contenus, via un lecteur ainsi qu'à une norme d'authentification des contenus avec d'autres acteurs comme la BBC, Adobe et Intel.

- **Le renforcement des contrôles préalables à l'utilisation de certains services** : Microsoft limite sur Bing certaines campagnes publicitaires liées à la COVID-19. L'opérateur limite par ailleurs les utilisations de son service *« d'IA cognitif »* Custom Neural Voice, afin de prévenir *« la prolifération des* deepfakes *»*.

Le Conseil constate néanmoins que des informations, notamment quantitatives, sur la performance de la détection et sur la viralité de ces contenus, manquent pour pouvoir évaluer l'efficacité de ces systèmes. Les opérateurs n'expliquent pas, ou ne le font que très partiellement, comment ces systèmes spécifiques fonctionnent, quelles sont leurs performances (faux positifs et négatifs) ainsi que leurs résultats concrets. Dans certaines déclarations, les informations spécifiques aux systèmes de modération sont éparses (Google, Snapchat, Twitter) voire inexistantes (Microsoft, LinkedIn, Verizon Media). **Il est donc impossible à ce stade de comprendre de quelle manière les opérateurs s'en saisissent et quels sont leurs effets sur la détection et la circulation des fausses informations.**

5.2.2. Le fonctionnement des algorithmes de recommandation et leur rôle dans la propagation de fausses informations

Les opérateurs abordent deux types d'utilisation de leurs systèmes algorithmiques de recommandation de contenus dans le cadre de leurs efforts pour limiter la propagation de fausses informations. Deux pratiques sont particulièrement évoquées dans les déclarations :

- **la mise en avant de sources fiables sur les services des plateformes** (Microsoft, Facebook, Google, Dailymotion, LinkedIn, Snapchat, YouTube) : cette approche adoptée par une grande majorité d'opérateurs consiste soit à rendre le service fermé pour tout (Dailymotion) ou partie (Snapchat) aux contenus amateurs, soit à faire remonter des résultats issus de sources réputées fiables en réponse à des requêtes d'utilisateurs sur les services (Bing, Google, LinkedIn, YouTube). Certains acteurs ont fait le choix d'organiser les informations réputées fiables au sein d'onglets ou de parties dédiées du service, comme des centres d'information vers lesquels les utilisateurs peuvent être redirigés ou des bandeaux d'information (Bing, Facebook, LinkedIn, YouTube) ;

- **la rétrogradation de contenus de faible qualité** (Microsoft, Dailymotion, Facebook) : Microsoft indique travailler à l'amélioration du classement et de la pertinence de ses résultats sur Bing et à *« limiter la visibilité des contenus de faible qualité et de faible autorité, comme les fausses informations »*. Dailymotion désindexe

les contenus signalés le temps de l'instruction ; (information confidentielle). Facebook indique utiliser notamment une méthode de *« réduction de la distribution des fausses informations et de la désinformation »*.

S'agissant du **fonctionnement des systèmes algorithmiques de recommandation de contenus en général** (et pas seulement en lien avec la lutte contre la manipulation de l'information), les déclarations de plusieurs opérateurs (Dailymotion, Facebook, Google, LinkedIn, Snapchat, Twitter, YouTube, la Fondation Wikimédia) comportent des éléments qui, dans l'ensemble, sont plus nombreux et de meilleure qualité que dans les déclarations précédentes. La ou les fonctionnalités du service auxquelles ils s'appliquent, leur objectif principal (indexation, classement, et affichage ou recommandation de contenus) et le ou les types d'algorithmes utilisés ont été plus précisément documentés. Pour ce dernier sujet, il s'agit principalement d'algorithmes d'apprentissage automatique (Google, Facebook) et de traitement automatique du langage naturel (Google, Wikipédia, Twitter).
Néanmoins, comme développé ci-après (partie 5.2.3), le Conseil constate que les éléments sur les systèmes algorithmiques liés à la recommandation de contenus demeurent trop généraux ou lacunaires pour lui permettre d'appréhender leur fonctionnement et les enjeux qui en découlent.

5.2.3. Une transparence globale accrue mais encore insuffisante

- **Des efforts à intensifier dans les informations communiquées aux utilisateurs sur les services**

Le Conseil a constaté une amélioration de la transparence vis-à-vis des utilisateurs, qui a gagné en précisions. Néanmoins, cette dernière ne concerne quasiment que les systèmes de recommandation, très peu d'informations étant fournies aux utilisateurs sur la modération algorithmique, qu'ils soient auteurs du signalement, auteurs du contenu ou qu'ils s'apprêtent à le consulter.

Les pratiques des opérateurs en matière de **transparence de la recommandation algorithmique** vis-à-vis de leurs utilisateurs se divisent en deux grandes catégories : celles consistant à fournir une information contextuelle sur un résultat donné de leurs systèmes algorithmiques (sur le service Bing pour les contenus publicitaires, Facebook pour les contenus publicitaires et organiques, Twitter pour les contenus modérés et considérés comme *« abusifs »*) et celles reposant sur la fourniture d'informations sur le fonctionnement général de ces systèmes au sein de rubriques dédiées (Dailymotion, Facebook, Google, LinkedIn, Twitter). La Fondation Wikimédia, qui a indiqué utiliser très peu d'algorithmes, met à disposition de ses bénévoles des outils permettant d'évaluer les contributions ainsi que les contributeurs pour lutter contre le vandalisme. Verizon Media – qui, pour son moteur de recherche Yahoo, renvoie à Bing – n'a fourni aucune information à ce sujet.

En termes de **paramétrages**, les utilisateurs sont informés qu'ils peuvent personnaliser la manière dont les contenus leur sont recommandés de manière proactive (Snapchat, lors de l'inscription sur la plateforme), ainsi que de manière contextuelle (à proximité des contenus organiques et publicitaires sur Facebook, dans les onglets Découvrir et Stories sur Snapchat) ou dans un espace d'aide (Bing, Facebook, Google, Snapchat, Twitter). Dailymotion ne permet pas de personnaliser les résultats de ses algorithmes de recommandation, tandis que LinkedIn et la Fondation Wikimédia ne déclarent pas de paramétrage spécifique possible. Verizon Media ne donne aucune indication à ce sujet.

Le CSA relève que Facebook propose désormais[32] à ses utilisateurs de choisir entre trois organisations de leur fil d'actualité : celle, « historique », qui résulte d'une recommandation algorithmique (onglet « Accueil ») ; une autre centrée sur les « amis » et « pages » choisis par l'utilisateur (onglet « Favoris ») ; une troisième répondant à une logique antéchronologique de présentation des contenus (« Récent »). Par défaut, à chaque connexion au service, l'utilisateur voit la version « Accueil ». Avec Twitter, qui déclarait l'année dernière proposer désormais, en plus de son fil d'actualité basé sur les dates de publication, une version faisant l'objet d'une recommandation algorithmique, il s'agit du second opérateur proposant plusieurs options d'organisation des contenus à ses utilisateurs. Le Conseil sera attentif lors du prochain exercice aux données concrètes d'usage et de retours utilisateurs qui seront fournies par ces opérateurs pour évaluer l'impact de ces choix d'organisation de fil d'actualité sur l'amélioration de la connaissance et de la compréhension par les utilisateurs de la recommandation algorithmique de contenus.

- **Les informations fournies au Conseil et au public dans les déclarations**

Webedia et Unify ont indiqué que cette section ne leur était pas applicable, car ils n'utilisent pas d'algorithmes pour organiser les contenus sur leur forum. Microsoft estime, sans explication ni précision sémantique, ne pas faire de recommandation de contenus, ce qui est assez surprenant concernant son service constitutif d'un moteur de recherche – d'autant plus que Verizon Media renvoie à Microsoft pour sa propre activité en la matière.

Dans les autres déclarations, la qualité et la quantité globales des informations transmises se sont notablement améliorées par rapport à l'exercice précédent. Les opérateurs ont décrit un certain nombre d'éléments de manière plus circonstanciée qu'en 2019. Quelques opérateurs ont donné des informations plus poussées sur les systèmes algorithmiques de recommandation et de modération qu'ils utilisent (Dailymotion, Facebook[33], Twitter, la Fondation Wikimédia, Snapchat[34]).

[32] Depuis mars 2021.
[33] Éléments recueillis lors de deux ateliers organisés par Facebook et qui avaient notamment pour objectif de répondre aux questions du CSA en la matière, les 3 et 10 juin 2021.
[34] À titre confidentiel.

Néanmoins, certains éléments ayant trait au fonctionnement des systèmes algorithmiques de recommandation comme de modération restent lacunaires et ne permettent pas une analyse exhaustive.

Les **types de données prises en entrée et en sortie des systèmes** algorithmiques de recommandation et de modération restent communiqués sous forme de listes d'exemples chez la quasi-totalité des opérateurs, sans que l'on ne puisse déterminer si elles sont exhaustives (Dailymotion[35], Facebook, Google, LinkedIn, Snapchat, Twitter). Chez certains, de telles listes ne sont tout simplement pas renseignées (Bing, Wikipédia, Yahoo Portal, Yahoo Search). Concernant la pondération de ces données, seul Facebook a apporté, oralement[36], des précisions concrètes.

Continuent par ailleurs de faire défaut chez une majorité d'opérateurs la **description complète des systèmes** de recommandation comme de modération reposant sur un ou plusieurs algorithmes (pas de listes des algorithmes sauf chez Dailymotion et la Fondation Wikimédia), les informations sur les **technologies utilisées** (hors Dailymotion, Facebook, Google, Twitter et la Fondation Wikimédia), le **niveau d'intervention humaine** (excepté Facebook, oralement, sur la modération, ainsi que Google, Snapchat et Twitter qui en font quelques mentions) ou les **grandes étapes de fonctionnement** de ces systèmes (hormis Facebook qui a déclaré les grandes étapes de fonctionnement de ses algorithmes liés à son fil d'actualité, et LinkedIn, dans sa déclaration pour l'année 2019). À l'exception de la déclaration de la Fondation Wikimédia[37], aucune **indication des performances** des systèmes utilisés n'a été indiquée.

Enfin, **très peu de changements structurants** ont été déclarés par rapport à l'exercice antérieur, ce qui est notable concernant une année lors de laquelle la circulation de nombreuses fausses informations en ligne a été documentée[38] à la faveur de la crise sanitaire et des élections municipales. Dailymotion, Facebook, LinkedIn et Snapchat ont indiqué ne pas avoir procédé à de tels changements dans leurs algorithmes de recommandation et de modération appliqués aux contenus consultés depuis la France en 2020. Microsoft, la Fondation Wikimédia et Verizon Media n'abordent pas le sujet. Seul Twitter a indiqué avoir davantage recouru à l'apprentissage automatique et à l'automatisation dans le cadre de la lutte contre les contenus potentiellement abusifs, avoir *« mis en place un système mondial de triage de la gravité du contenu »* et avoir *« effectu[é] des contrôles quotidiens de qualité sur nos processus d'application sur les contenus »* en 2020.

[35] Des précisions ont été données, mais à titre confidentiel.
[36] Lors de deux ateliers les 3 et 10 juin 2021.
[37] L'un des algorithmes développés pour détecter le vandalisme, ClueBot NG, a un taux de performance de 90 %, taux de classement correct des contributions comme étant « *du vandalisme ou non-vandalisme* ».
[38] Voir partie 4 sur l'impact de la crise de la Covid-19 (p.17).

Il convient de reconnaître la complexité de rendre compte de ces éléments s'agissant de certaines technologies, notamment l'apprentissage automatique auxquels ont recours Facebook, Google, Snapchat, Twitter, la Fondation Wikimédia ou YouTube. Néanmoins, en l'état des déclarations, le CSA n'est pas en mesure d'évaluer l'efficacité des mesures prises par les opérateurs pour assurer une meilleure transparence de leurs systèmes algorithmiques vis-à-vis des utilisateurs et du régulateur. Pour cela, des informations précises sur les effets de ces systèmes et sur ce à quoi accèdent les utilisateurs auraient été nécessaires. Il aurait en outre fallu que les déclarations soient plus explicites sur le caractère exhaustif ou non des éléments déclarés, concernant notamment les grandes étapes de fonctionnement de ces systèmes, celles sur lesquelles les utilisateurs peuvent influer, les catégories de données prises en entrée ou encore les éventuelles pondérations appliquées et leurs modalités.

De manière générale, la fourniture, par les opérateurs, de métriques permettant de caractériser les mesures de transparence qu'ils mettent en œuvre permettrait de renforcer cette transparence. Ainsi, **les opérateurs n'ont pas fourni la liste demandée par le Conseil des contenus publicitaires et organiques détectés comme contenant une fausse information ayant suscité le plus d'interaction de la part des utilisateurs en 2020**.

Enfin, le niveau d'informations reste très similaire à celui en accès libre. Deux déclarations ont fait l'objet de mentions de confidentialité : Dailymotion sur ses systèmes de recommandation et Snapchat sur ses systèmes dédiés à la lutte contre la manipulation de l'information. À cet égard, le CSA rappelle de nouveau aux opérateurs qu'ils ont la possibilité de marquer, dans leur déclaration, les éléments couverts par le secret des affaires afin qu'ils soient traités de manière strictement confidentielle. Il les appelle à se saisir de cette possibilité pour lui fournir toutes les informations lui permettant d'évaluer les mesures prises en matière de transparence des algorithmes.

5.2.4. La prise en compte par un nombre restreint d'opérateurs des notions de loyauté, d'explicabilité et d'équité

Le CSA relève que plusieurs opérateurs ont abordé des problématiques éthiques liées à la mise en œuvre de leurs systèmes algorithmiques.

Dans son questionnaire, le CSA a proposé des définitions des notions de loyauté, d'explicabilité et d'équité, formulées avec l'appui de son comité d'experts sur la désinformation, et a appelé les opérateurs à les commenter ainsi qu'à indiquer comment ils les mettaient en œuvre. Il salue les réponses de LinkedIn, Microsoft, Twitter, Dailymotion et Facebook et estime que les autres répondants auraient également dû se saisir de cette

possibilité d'approfondir le dialogue avec le régulateur et la société civile sur la question de la transparence des algorithmes.

L'équité est le principe qui a été le plus évoqué. LinkedIn a expliqué se concentrer en particulier sur cette notion et avoir *« développé et mis en libre accès »* une boîte à outils afin que d'autres entreprises l'utilisent notamment pour *« mesurer les biais dans les données de formation »*. Microsoft lie la mise en œuvre de ce principe sur Bing à sa politique de lutte contre les biais ainsi qu'aux six principes qu'il suit pour *« guider le développement de l'intelligence artificielle »* que sont *« l'équité, la fiabilité et la sécurité, la vie privée et la sécurité, l'inclusion, la transparence et la responsabilité »*. Il mentionne l'existence d'une équipe en interne ainsi qu'un conseil consultatif travaillant sur ces sujets. Twitter a indiqué au contraire ne pas utiliser le concept d'équité car *« il implique un jugement moral »* et lui préfère le terme de biais. Il déclare évaluer ses systèmes utilisant de l'apprentissage automatique afin de vérifier qu'ils *« produisent des informations et des recommandations qui sont transparentes, responsables et évitent autant que possible les biais »*. Facebook s'est attaché plus particulièrement à la notion de loyauté, définie comme un *« processus »* plutôt qu'une *« propriété d'un système »*. L'opérateur indique mener un travail spécifique sur le sujet et a des équipes dédiées en interne afin de développer une *« intelligence artificielle fiable »* et des *« produits justes et équitables »*.

Dailymotion estime que ces définitions n'ont pas vocation à s'appliquer à son service car *« ces critères sont pertinents s'ils s'appliquent à des réseaux sociaux qui mettent en valeur des contenus amateurs et dont l'audience est essentiellement* onsite *»*.

Plusieurs opérateurs ont détaillé leurs actions de lutte contre les biais dans les données d'apprentissage des algorithmes d'apprentissage automatique qu'ils utilisent. Certains ont ainsi fait référence au développement d'outils ou d'équipes internes mais avec des niveaux de précision disparates. LinkedIn mentionne un outil développé et disponible en *open source* à destination d'autres entreprises. Microsoft a élaboré six principes, des *« bonnes pratiques »* et donne un exemple concret d'application à son outil Custom Neural Voice. Snapchat fait référence à certaines procédures d'évaluation de l'équité de ses classificateurs menées par son équipe de recherche interne et à l'implication de modérateurs humains dans certaines tâches afin d' *« assurer que l'automatisation ne soit pas déconnectée de nos politiques ni des nuances inhérentes à la modération humaine »*. Twitter mentionne le travail de son équipe interne META[39] qui *« utilise des méthodologies et des cadres d'investigation bien établis pour comprendre et atténuer la partialité de nos produits alimentés par le* machine learning *»*. Dailymotion explique que les spécificités de son service le prémunissent de tout biais car il en exclut *« systématiquement les contenus amateurs »* et *« ne recourt à aucun algorithme dans la prise de décision relative à la modération des contenus de désinformation »*.

[39] Machine Learning Ethics, Transparency and Accountability.

Le CSA estime qu'il aurait essentiel que tous les opérateurs ayant déclaré utiliser l'apprentissage automatique évoquent le sujet central de la lutte contre les biais (notamment Google, la Fondation Wikimédia).

Au vu de ces éléments d'analyse, le CSA formule les préconisations suivantes :

- **Proposer aux utilisateurs des fonctionnalités leur permettant de comprendre**, si possible de manière personnalisée et contextuelle, **les effets des systèmes algorithmiques de recommandation et de modération**.
- **Déclarer tout élément donnant à voir comment la lutte contre les biais est concrètement mise en œuvre sur leurs services** : ressources dédiées, outils, modifications opérées par la suite, résultats.

5.3. Promotion des contenus issus d'entreprises et d'agences de presse et de services de communication audiovisuelle

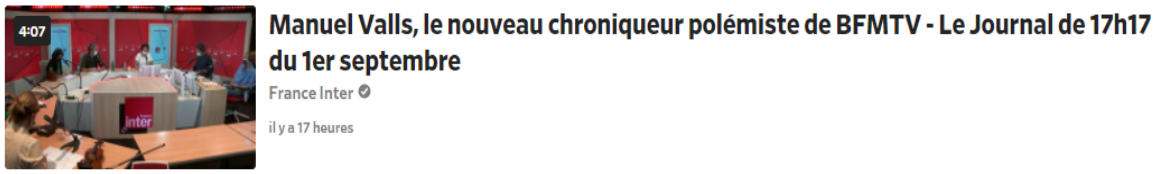
En 2020, les mesures en faveur de la diffusion d'une information vérifiée sont globalement, identiques à celles déclarées pour l'exercice précédent. Dès lors, **la majeure partie de l'analyse et des conclusions formulées dans le précédent bilan du Conseil restent d'actualité.**

Trois plateformes (Doctissimo, Jeuxvideo.com et Wikipédia) continuent d'indiquer ne pas prendre de mesures en la matière en raison de la nature même de leur service : les deux premières, parce que dans les faits, les personnes qui utilisent le service sont des utilisateurs privés et non des organismes (notamment de presse), la troisième parce qu'elle place le principe de sources fiables au cœur de son fonctionnement.

Deux autres opérateurs déclarent, comme l'an dernier, sélectionner en amont les partenaires médias dont les contenus sont disponibles au sein de tout ou partie de leur service (Snapchat (dans la partie Discover de son service) et Verizon Media) et estiment moins nécessaire, de ce fait, le recours à la vérification des informations publiées. Deux autres ne déclarent aucune nouvelle mesure en la matière (Dailymotion et LinkedIn).

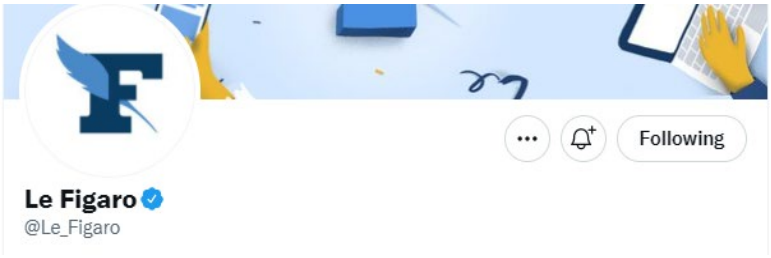
Les nouvelles mesures déclarées ont souvent été mises en œuvre dans le cadre du contexte particulier de l'année 2020 (Covid-19 ou élection présidentielle américaine) et n'ont pas nécessairement été systématisées. Elles portent surtout sur l'identification des sources de contenus.

Labellisation/certification par Dailymotion[40] :



Labellisation/certification par Facebook[41] :

Labellisation/certification par Twitter[42] :



5.3.1. Les nouvelles mesures relatives à l'identification des sources de contenus.

Les nouvelles mesures déclarées sont mises en œuvre dans le but de donner davantage d'éléments de contexte aux utilisateurs, notamment en libellant plus systématiquement certains types de contenus (Facebook, Google et Twitter).

Facebook et Twitter libellent dorénavant les contenus issus de médias affilés à un État ou sous son contrôle ; cette mesure contribue à éclairer les utilisateurs en identifiant une éventuelle propagande relayée par les « médias d'État ». Twitter déclare libeller les tweets susceptibles d'inclure un contenu média modifié. Google annonce l'apparition d'un libellé

[40] Source : captures d'écran réalisées en août 2021.
[41] Source : observations CSA, août 2021.
[42] Source : observations CSA, août 2021.

« *vérification des faits* » au sein des parties *Actualités* et *Images* de son moteur de recherche afin de signaler les articles comportant un travail de *fact-checking*.

Libellés Twitter sur les contenus ou comptes affiliés à un Etat[43] :

Facebook a mis en place un index des pages publiant principalement du contenu d'actualité. Outil à usage des éditeurs et sans incidence sur l'expérience utilisateur, cette indexation se fait à la demande des éditeurs qui doivent respecter des critères élaborés avec des organismes de presse, des universitaires et organisations industrielles. Facebook peut vérifier les informations renseignées par les éditeurs sur leur identité. Cette mesure tend à répondre à la recommandation adressée par le Conseil aux plateformes dans son bilan précédent, d'œuvrer à la vérification de l'exactitude des informations apportées par les entreprises et agences de presse et services de communication audiovisuelle lorsque qu'ils s'identifient eux-mêmes sur les services des plateformes.
Facebook et Google donnent davantage d'informations sur leur mécanisme d'identification des contenus, notamment ceux relevant de l'actualité, par des modèles d'apprentissage automatique alimentés par des modérateurs (Facebook) ou des évaluateurs externes (YouTube).

Microsoft a créé des outils destinés aux éditeurs de contenus, sous forme de solutions techniques d'identification ou d'authentification des contenus (*Microsoft Video Authenticator*) et de certificats numériques.

Twitter fait état de nouvelles mesures déployées dans le cadre de l'élection présidentielle américaine. En association avec l'organisation *Ballotpedia*, la plateforme a libellé les comptes, les tweets électoraux et les retweets de candidats en y associant des informations de contexte. Les contenus porteurs de fausses informations étaient soit retirés, soit étiquetés comme trompeurs et manipulateurs et, en cas de repartage, une fenêtre avertissait l'utilisateur qui souhaitait partager le contenu que ce dernier était faux et trompeur. Une telle mesure est de nature à éclairer les utilisateurs et à concourir à limiter la viralité des fausses informations, notamment dans une période de menaces accrues envers la sincérité d'un scrutin. Il serait néanmoins intéressant de connaître la qualification des rédacteurs de

[43] Source : observations CSA, août 2021.
[44] Source : observations CSA, août 2021.

ces informations et la part du recours au travail journalistique pour vérifier ou étayer ces informations.

5.3.2. Les nouvelles mesures relatives à la prise en compte des démarches de labellisation

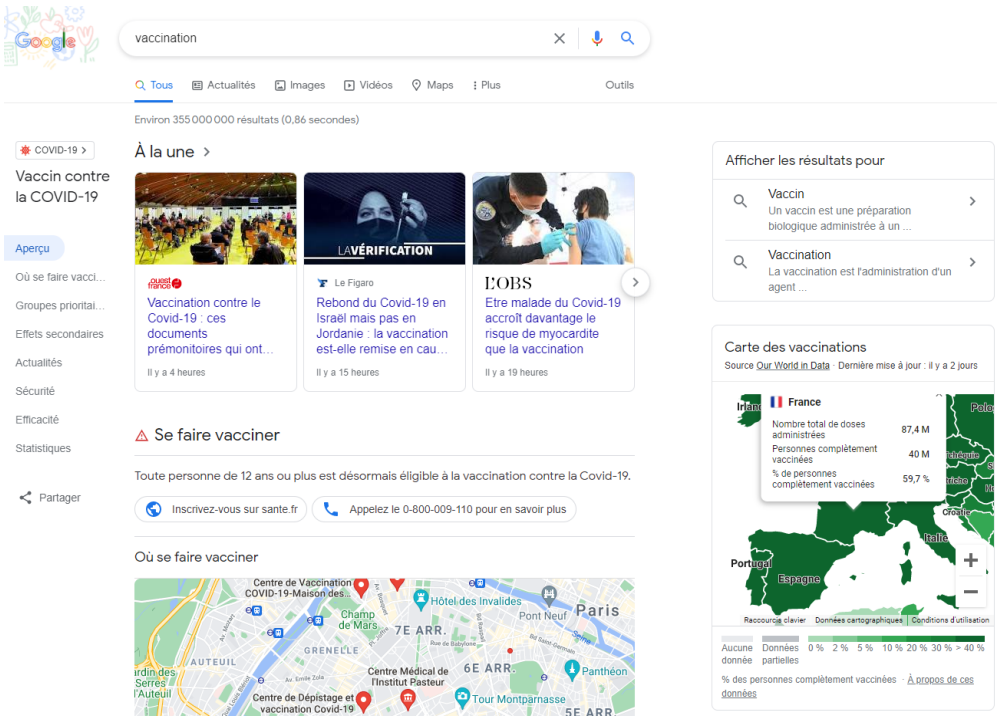
Les nouvelles mesures déclarées en matière de prise en compte des démarches de labellisation prennent la forme de création de partenariats et de participation, notamment financière, à des programmes visant à promouvoir un journalisme de qualité ou à élaborer des normes. Certains programmes sont à l'initiative des plateformes.

En revanche**, peu de nouveaux éléments sont apportés sur la recherche et la prise en compte des démarches de labellisation par les plateformes**, notamment dans l'identification des sources (comme le préconise la recommandation du Conseil).

Facebook a renforcé ses collaborations avec les gouvernements et les experts, ainsi qu'avec l'*International Fact-Checking Network* (IFCN). Microsoft a créé un organisme, intitulé *Coalition for Content Provenance and Authenticity*, dont le but est d'élaborer des normes ouvertes et des spécifications techniques sur la provenance et l'authentification des contenus.

À travers son programme *Google News Initiative*, Google a financé, dans le contexte de la pandémie de Covid-19, un hub médiatique afin de regrouper les informations relatives à la vaccination et faciliter le travail de vérification de l'information. Dans le cadre de l'élection présidentielle américaine, Twitter a également créé un hub réunissant les informations relatives à l'élection et a délivré des formations aux équipes de campagne. Il aurait été intéressant d'avoir des informations détaillées sur les critères utilisés pour sélectionner les informations reprises dans ces hubs et sur la façon dont les opérateurs intégraient le travail de « *fact-checking* » des journalistes.

Hub médiatique de Google[45] **:**



Enfin, Snapchat a déclaré des informations confidentielles à ce sujet.

5.3.3. Les nouvelles mesures relatives à la mise en avant des contenus issus des entreprises et agences de presse et services de communication audiovisuelle, notamment de *fact-checking*

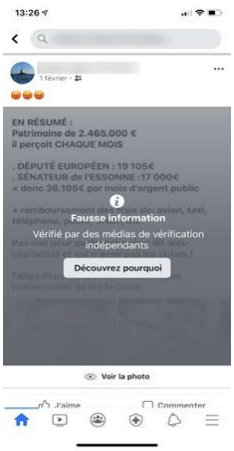
Parmi les nouveautés significatives, on note la création par Google de l'outil *Fact Check Explorer*, moteur de recherche accessible à tous recensant les articles de presse vérifiant la véracité d'une information en lien avec le mot-clé saisi. La mise en place d'un tel outil doit être saluée en ce qu'il permet à tout utilisateur d'obtenir une information vérifiée et permet de faire connaître le travail journalistique.

[45] Source : observations CSA, août 2021.

Mention de *Fact-checking* sur Google Search[46] :

Les mécanismes mis en œuvre par Facebook en la matière sont particulièrement avancés et permettent de donner davantage d'informations fiables aux utilisateurs. Facebook a complété les moyens déjà déployés en mettant à disposition de ses partenaires tiers vérificateurs deux nouvelles options d'évaluation des contenus – contenu modifié et contexte manquant – afin d'affiner l'instruction du contenu et l'information à destination des utilisateurs. Facebook apporte utilement des informations sur le fonctionnement de la procédure de saisine de ces partenaires sur les contenus à vérifier, mêlant intelligence humaine et technologique, avec auto-saisine possible des partenaires, dont les évaluations viennent alimenter le modèle d'apprentissage automatique.

Mention de *fact-checking* sur Facebook et Instagram[47] :



Bien qu'il soit compréhensible que les informations relatives au budget et aux accords financiers entre la plateforme et les tiers vérificateurs soient confidentielles, de telles informations seraient de nature à éclairer l'analyse du Conseil et lui permettraient de préciser ses futures recommandations. Il est également compréhensible que le système de remontée et de mise à disposition des fausses informations aux tiers vérificateurs, s'appuyant sur différents signaux dont une partie automatisée et permettant une saisine par les tiers vérificateurs eux-mêmes, ne facilite pas la tenue de liste et le recensement du nombre de contenus véhiculant de fausses informations et le nombre de contenus vérifiés.

[46] Source : déclaration de l'opérateur.
[47] Source : déclaration de l'opérateur.

Néanmoins, de telles informations seraient utiles à l'analyse du Conseil et des lecteurs du présent bilan.

Concernant Twitter, son recours aux tiers vérificateurs et la délivrance d'informations contextuelles quant à l'identité des sources et à la fiabilité de certains contenus ne sont ni larges ni systématiques. Cette situation peut paraître paradoxale pour une plateforme qui déclare ne pas vouloir être l'arbitre de la vérité.

Au vu de ces éléments d'analyse, le CSA formule les préconisations suivantes :

- **Conserver une part d'intervention humaine dans la vérification des informations fournies au sujet des organes de presse lorsqu'ils font l'objet d'une identification particulière**.

- **Développer des initiatives et partenariats tels que ceux pris dans le contexte de la crise sanitaire** (en coopération avec des pouvoirs publics, des associations, des chercheurs, etc.) tout en veillant, le cas échéant, à distinguer les sources gouvernementales, à des fins de transparence vis-à-vis des utilisateurs.

5.4. Lutte contre les comptes propageant massivement de fausses informations

Le législateur invite les plateformes à prendre de mesures pour lutter contre les « *comptes propageant massivement de fausses informations* ». Cette notion ne vise pas un type de contenus, mais des pratiques développées par des utilisateurs de la plateforme afin de diffuser massivement des fausses informations susceptibles de troubler l'ordre public ou d'altérer la sincérité d'un scrutin. Elle est, par nature, **protéiforme, ces pratiques pouvant ainsi différer en fonction des caractéristiques de la plateforme, se recouper et ne pas être intrinsèquement liées à la diffusion de fausses informations.**

Dailymotion, Microsoft, Snapchat, Unify et Verizon Media ont indiqué ne avoir eu connaissance de ce type d'usages sur leurs services respectifs.

5.4.1. La notion de « comptes propageant massivement de fausses informations »

Au vu des déclarations de l'année passée, le CSA a interrogé les opérateurs, dans son questionnaire concernant l'exercice 2020, sur leur appréhension de la notion afin de faire un

état des lieux, au vu des éléments déclarés, des types de pratiques et de comptes concernés et des moyens de lutte spécifiquement déployés pour les contrer.

Les réponses des opérateurs permettent de proposer la liste de pratiques qui suit, non exhaustive ni figée. Il convient de noter que les catégories identifiées peuvent se recouper (ex. : le spam peut être un des moyens d'une campagne d'influence coordonnée) et qu'elles ne sont pas nécessairement liées à un usage malveillant de fonctionnalités permettant de rendre viraux des contenus.

- **Les faux comptes et comptes trompeurs**

L'attention des opérateurs n'est pas uniquement portée sur les comptes diffusant massivement de fausses informations mais également sur les comptes trompeurs (faux comptes) de manière générale, qui sont à l'origine d'une grande partie de contenus nuisibles sur les services, ces contenus ne se limitant pas à la manipulation de l'information.

On note que ce type de comptes ne participe pas nécessairement, individuellement, d'une massification de la diffusion d'une fausse information mais qu'ils sont particulièrement susceptibles d'être créés à cette fin.

- **Les pratiques d'influence coordonnées**

Trois opérateurs ont fourni des éléments étayés, avec des exemples précis, permettant de mieux appréhender la notion.

Facebook définit les campagnes d'influence comme des efforts coordonnés pour manipuler le débat public dans un but stratégique, où les faux comptes sont au cœur de l'opération. Son approche des campagnes d'influences coordonnées se fonde sur le comportement et les acteurs : au lieu d'identifier les contenus violant les politiques internes (telles que la désinformation et les discours haineux), l'opérateur identifie les comportements qui sont en contradiction avec les règles spécifiques en matière de comportement inauthentique. Il est intéressant de noter que Facebook distingue deux catégories : les comportements inauthentiques coordonnés (campagnes nationales non gouvernementales) et les ingérences étrangères ou gouvernementales (campagnes d'influence menées au nom d'une entité gouvernementale ou par un acteur étranger).

Google fait état d'opérations d'influence coordonnées, Twitter d'activités inauthentiques coordonnées.

- **L'utilisation de technologies avancées (IA) pour créer des contenus trompeurs : cas des *deepfakes***

La définition de cette pratique diffère légèrement d'une plateforme à une autre : « *vidéos manipulées et trompeuses* » (Facebook**)** ; « *médias synthétiques et manipulés* (Twitter). Microsoft estime que les *deepfakes* présentent de nombreuses utilisations bénignes et des avantages potentiels, mais qu'ils peuvent dans le même temps être utilisés pour porter atteinte aux réputations et saper la confiance des institutions démocratiques. Il convient en effet de rappeler que les contenus de types *deepfakes* ne sont pas, en soit, nécessairement problématiques.

- **Pratiques trompeuses massives non nécessairement liées à une tentative de nuire à l'ordre public ou à la sincérité d'une élection**

La propagation massive de fausses informations peut se faire via des pratiques globales massives qui ne lui sont pas spécifiques, pouvant également servir des fins purement lucratives par exemple. C'est le cas du spam qui peut être utilisé dans des contextes électoraux pour influencer sur les votes. C'est également le cas de l'engagement artificiel et trompeur (ex. : gonflement des « *likes* » ou du nombre d'abonnés).

- **Pratiques de « flood » et de « up » sur les forums**

Webedia considère que par sa nature, Jeuxvideo.com ne permet pas le relai massif de fausses informations : les contenus publiés sur le forum sont accessibles à l'ensemble des internautes et affichés de manière chronologique ou antéchronologique. On note néanmoins qu'à l'échelle d'un forum, les pratiques de *flood* (multi-publication d'un contenu) et de *up* (publication de messages intempestifs dans un fil de discussion pour le faire remonter dans la liste des fils de discussion) constituent des techniques visant à amplifier artificiellement la visibilité d'un contenu et sont donc susceptibles d'être utilisées par un utilisateur malveillant souhaitant propager largement une fausse information.

- **Opérateurs estimant ne pas être soumis au risque d'une propagation massive de fausses informations en raison des spécificités de leur service**

Certains opérateurs considèrent que la nature de leur service empêche le relai massif de fausses informations (Dailymotion, Snapchat). À ce titre, Snapchat rappelle que les discussions de groupe sont limitées à 64 « amis » et qu'un grand nombre de fonctionnalités du service sont, par défaut, paramétrées comme privées, ce qui protègerait les utilisateurs contre le partage involontaire d'informations. Il n'en reste pas moins que certaines parties du service (la « Map », notamment) donnent accès à tous les utilisateurs à des contenus publics.

5.4.2. Les moyens de détection et les mesures de lutte contre ces pratiques

- **Moyens déployés**

La majorité des opérateurs englobent la lutte contre la viralité des contenus contenant de fausses informations dans celle contre l'ensemble des contenus violant leurs politiques internes, notamment s'agissant de la détection et du traitement des spams avant ou après leur création (Facebook, Google, qui lutte contre les tentatives de manipulation artificielle du taux d'engagement telles que la mention « *je n'aime pas* » sur les vidéos, pratique interdite par les règles internes).

Concernant les moyens humains et financiers, la grande majorité des opérateurs recourent à la même équipe de modération que celle dédiée au traitement des signalements de contenus et le budget affecté au traitement des comptes se recoupe avec celui du traitement des signalements (Dailymotion, Facebook).

Concernant l'identification et le traitement des opérations d'influence coordonnées, les informations apportées sont plus spécifiques puisque le travail est mené soit par une équipe dédiée, soit par plusieurs équipes dont l'une peut être dédiée (LinkedIn).

- **Mesures de détection et de modération**

La modération diffère en fonction du type de pratiques considérées : les pratiques individuelles sont traitées dans le cadre des moyens de modération généraux, tandis qu'il existe des actions plus spécifiques sur les pratiques coordonnées (avec des équipes dédiées à la conduite d'enquêtes).

S'agissant des spams, les clôtures de compte résultent, dans la majorité des cas, d'un autre motif que la manipulation de l'information, la pratique étant généralement interdite indépendamment du but poursuivi par son auteur**.**

De même, les comptes pratiquant l'usurpation d'identité et cherchant à propager de fausses informations en se présentant de manière inexacte sont supprimés en ce qu'ils violent les politiques internes des plateformes (Google).

De manière générale, les moyens de lutte contre les comptes et pratiques de propagation massive de fausses informations combinent moyens humains et automatisés, dans des proportions variables en fonction de la pratique concernée[48]. Il ressort des déclarations que les opérateurs sont confrontés à la nécessité de faire évoluer constamment ces moyens de lutte, notamment technologiques, à mesure que les acteurs malveillants apprennent à les contrer.

Concernant les *deepfakes*, les opérateurs adoptent des mesures de lutte différentes selon qu'ils considèrent qu'il s'agit d'une pratique bien établie (développement d'outils de détection) ou plutôt naissante (détection et actions prises sur les hypertrucages).

Enfin, le Conseil salue la mise en place de coopérations entre les plateformes et d'autres organismes, notamment avec le monde de la recherche, pour lutter contre les modèles et les pratiques caractéristiques de ces phénomènes (Facebook, LinkedIn, Microsoft et Twitter).

- **Mesures spécifiques à l'encontre des comptes liés à des opérations d'influence coordonnées**

Les mesures prises par Facebook à l'encontre des comportements inauthentiques coordonnés sont spécifiques : suppression des comptes, pages et groupes impliqués dans ces opérations. Facebook opère une veille des réseaux précédemment supprimés et échange des informations avec des chercheurs indépendants, des organismes gouvernementaux et des partenaires industriels.

Google considère que les campagnes d'influence sont moins fréquentes sur ses plateformes que sur les autres en raison de l'impossibilité de partager des informations directement entre utilisateurs. Il détaille son approche opérationnelle et les mesures prises à l'encontre des pratiques coordonnées en présence de tentatives de piratage et d'hameçonnage et de comptes liés à des opérations d'influence coordonnées (étrangères et nationales) identifiées sur ses services. Là aussi, des mesures spécifiques sont déployées (suppression de chaînes YouTube et blogs des acteurs impliqués, option de monétisation réduite).

- **Mesures d'information des utilisateurs**

Pour certains opérateurs, l'information des utilisateurs sur les mesures prises à l'encontre des comptes et pratiques de propagation massive et des risques encourus concerne indifféremment l'ensemble des types de contenus prohibés par les règles internes – dont les fausses informations (c'est le cas de Dailymotion).

[48] Voir en annexe.

Pour d'autres, les règles appliquées concernent les règles relatives à l'intégrité et l'authenticité d'un compte, qui peuvent comprendre les comportements trompeurs, la cybersécurité et les comptes inauthentiques (Facebook).

Concernant les opérations d'influence coordonnées, Facebook publie des rapports mensuels sur les comportements inauthentiques coordonnés[49] comprenant des données sur les comptes, pages et groupes faisant l'objet d'une action et détaillant les raisons pour lesquelles des réseaux sont démantelés ainsi que les dispositifs déployés selon le pays concerné.[50] Google publie un bulletin trimestriel sur les mesures prises à l'encontre des comptes qu'il lie à des campagnes d'influence coordonnées étrangères et nationales[51]. Enfin, Twitter documente les comptes liés à des opérations de manipulation de l'information soutenues par des Etats dans des archives publiques alimentées depuis octobre 2018.[52]

Au vu de ces éléments d'analyse, le CSA formule les préconisations suivantes :

- **Informer davantage les utilisateurs sur les pratiques coordonnées d'influence et les risques qui en découlent,** tout particulièrement en période électorale.
- **Faire preuve de davantage de transparence, à tout le moins à l'égard du régulateur, sur les recettes publicitaires potentielles générées par les comptes propageant massivement de fausses informations.**

[49] Accessibles au lien suivant : https://about.fb.com/news/tag/coordinated-inauthentic-behavior/.
[50] En France, en décembre 2020, 84 comptes Facebook, 14 comptes Instagram, 6 pages et 9 groupes initiés en France ont été supprimés.
[51] En décembre 2020, Google a notamment supprimé trois chaînes YouTube dans le cadre d'opérations d'influence liées à la France critiquant le gouvernement russe et visant la République centrafricaine et le Mali (campagne également détectée par Facebook) et 3317 chaînes YouTube liées à la Chine.
[52] En juin 2020, l'opérateur a divulgué 32 242 comptes liés à trois opérations distinctes qu'il a attribuées à la République populaire de Chine, à la Russie et à la Turquie. Ces divulgations visent à protéger l'intégrité de la conversation publique et à informer les tiers enquêteurs (public, journalistes, chercheurs).

5.5. Mesures de lutte contre les fausses informations en matière de communications commerciales et de promotion des contenus d'information se rattachant à un débat d'intérêt général

La recommandation du Conseil de 2019 formule des orientations à l'attention des plateformes en matière d'information des utilisateurs sur la nature, l'origine, les modalités de diffusion des contenus et l'identité des personnes versant des rémunérations en contrepartie de la promotion des contenus d'information.

Les travaux menés par le Conseil et l'examen des déclarations concernant l'exercice 2019 ont fait apparaître que l'un des points cruciaux de la lutte contre la diffusion des fausses informations résidait, plus largement, dans les **liens entre flux financiers et fausses informations.** En effet, via la publicité et les fonctionnalités de monétisation des contenus, ces dernières peuvent, d'une part, être source de revenus pour différents types d'acteurs et, d'autre part, bénéficier, directement ou non, des mécanismes de diffusion des contenus commerciaux. En conséquence, ces liens peuvent être un facteur d'augmentation de la viralité des fausses informations et d'incitation à les diffuser. C'est pourquoi, dans le questionnaire adressé aux opérateurs en janvier 2021, le Conseil a choisi d'interroger les opérateurs sur l'ampleur réelle des liens entre communications commerciales et diffusion de fausses informations ainsi que sur les moyens de lutte mis en œuvre en la matière.

De manière générale, le niveau d'information fourni par les répondants est globalement élevé et les détails sont plus nombreux que lors du précédent exercice, notamment en ce qui concerne les mesures adoptées pour garantir la sécurité des marques et des annonceurs. Néanmoins, certains manques subsistent et, cette année encore, le niveau des informations communiquées varie grandement d'une plateforme à l'autre.

5.5.1. Un besoin de clarification terminologique afin de couvrir la diversité des pratiques et des modèles

En raison de la diversité des services de plateformes et des fonctionnalités qu'elles offrent, les communications commerciales qu'elles hébergent sont de natures variées et relèvent de différentes pratiques. **Le Conseil a donc proposé de considérer trois catégories de communications commerciales :** l'annonce publicitaire, le contenu sponsorisé et le contenu d'utilisateur en partenariat.

Il a invité les opérateurs à lui faire part de leur appréciation et remarques sur ces catégories, afin de les ajuster le cas échéant pour qu'elles couvrent au plus près les pratiques existantes.

Trois opérateurs ont réagi : Facebook partage cette typologie, qui correspond à ses pratiques, ainsi que Dailymotion, qui précise utilement en quoi les différentes catégories

sont pertinentes ou non au regard de son modèle. LinkedIn fait part d'un type particulier de communications commerciales, les « *campagnes InMail* sponsorisées », par messages dans la boîte de réception. **Les autres plateformes n'ont fait aucune remarque** sur cette proposition de catégorisation, ce que le Conseil estime qu'elles auraient dû faire. Outre le dialogue constructif avec le régulateur, cela aurait contribué à une meilleure transparence sur le fonctionnement et l'activité de services dont **le modèle économique repose essentiellement sur la publicité**.

5.5.2. Un besoin de transparence accrue pour les utilisateurs

L'ensemble des plateformes disposent de politiques publicitaires publiques, dont le Conseil estime qu'il est essentiel qu'elles soient facilement accessibles, compréhensibles et disponibles en français.

Sur les services, les communications commerciales sont distinguées des autres contenus par une labellisation visuelle particulière. Plusieurs opérateurs *(*Facebook, Google, Snapchat, Twitter, Microsoft *dans certains cas)* donnent quelques indications à l'utilisateur sur le ciblage commercial[53]. Peu de nouveautés ont été signalées par les plateformes sur ce point. Il convient toutefois de relever la mise en place d'un nouvel outil par Google permettant à l'utilisateur d'obtenir davantage d'informations sur un annonceur et de manière plus aisée. De la même manière, LinkedIn a développé une fonctionnalité permettant à ses membres d'obtenir plus d'éléments sur le ciblage des communications commerciales. Les déclarations de Facebook, LinkedIn et Twitter comportent des explications détaillées sur le fonctionnement des espaces publicitaires et du ciblage.

Certaines plateformes permettent à l'utilisateur de paramétrer ses préférences publicitaires, dans une logique de transparence et de responsabilisation. Il est toutefois dommage que ces outils n'expliquent pas pourquoi la plateforme associe l'utilisateur à telles ou telles catégories de ciblage. Le CSA estime à cet égard que les plateformes **ne donnent pas davantage d'information, dans leur déclaration, sur ce que recouvre ce paramétrage.**

Enfin, le Conseil invite les plateformes à **intensifier leurs efforts concernant le suivi des signalements des utilisateurs relatifs aux annonces publicitaires et aux contenus promus.**

[53] Voir, à cet égard, les outils contextuels « Pourquoi je vois cette publicité ? » sur Facebook, « Pourquoi est-ce que je vois cette publicité ? » sur LinkedIn et « Pourquoi cette annonce ? » sur YouTube.

5.5.3. Les mesures en faveur de la sécurité des marques (« *brand safety* »)

Les mesures de sécurité des marques permettent de protéger l'image des annonceurs, dont il n'est pas dans l'intérêt de voir leur marque associée à un contenu véhiculant de fausses informations. De plus, prévenir de telles situations permet d'empêcher la monétisation de fausses informations et, ainsi, de lutter contre leur développement. Si certaines plateformes ont donné davantage d'informations que l'année passée sur les mesures mises en place en la matière (Facebook), d'autres répondants restent encore imprécis (Snapchat, Verizon Media, Microsoft, Google).

Les mesures varient d'un opérateur à l'autre. On note la mise en place d'équipes internes ou/et de partenariats avec des tiers indépendants (Twitter, Facebook, Dailymotion). Les moyens déployés peuvent relever du design de la plateforme (interdiction de publicités superposées ou en *pop-up*), de ses actions de modération, de processus de validation préalable des annonceurs (Google, Twitter), de moyens donnés aux annonceurs en amont (possibilité de sélectionner les placements : Facebook, Dailymotion ; possibilité de travailler avec des tiers de confiance pour sélectionner les espaces adaptés à la campagne : Facebook) et en aval (rapport de diffusion : Twitter, Facebook ; voies de recours et de compensation : Dailymotion). **Le CSA encourage ces mesures, notamment celles visant à responsabiliser les annonceurs. Elles peuvent contribuer, à terme, à limiter la monétisation des fausses informations.**

En revanche, la majorité des plateformes ne donne pas d'indications chiffrées sur ces moyens ni sur l'impact concret des mesures prises en matière de sécurité des marques. Le CSA les invite à partager à l'avenir davantage d'informations sur ces points.

5.5.4. La (dé)monétisation des contenus organiques et publicitaires véhiculant de fausses informations : manque d'information sur l'ampleur du phénomène

Les mesures de sécurité des marques et des espaces publicitaires répondent à des risques réels dont il **serait nécessaire de mieux identifier l'ampleur**. Or, les plateformes ne fournissent pas d'information chiffrée sur le volume d'annonces publicitaires et de contenus sponsorisés porteurs de fausses informations ou accolés à un contenu véhiculant une fausse information. De la même manière, elles ne mentionnent ni les revenus générés par ces publicités avant d'être repérées ni le nombre d'utilisateurs exposés à de tels contenus avant leur retrait. Quelques données intéressantes sont communiquées (Google, Microsoft) mais elles restent générales et à une échelle mondiale.

Des informations plus précises et à l'échelle locale sont nécessaires pour comprendre l'ampleur du phénomène. Le Conseil appelle l'ensemble des opérateurs à faire preuve de davantage de coopération et de transparence sur ce point à l'avenir.

Les raisons de ce silence, lorsqu'elles sont fournies, sont de plusieurs natures. Certains répondants considèrent qu'il est trop complexe de fournir de telles données. D'autres ne s'estiment pas concernés car les contenus organiques ou publicitaires porteurs de fausses informations n'existeraient pas sur leur plateforme. Certains estiment que les politiques mises en place (détection et blocage automatiques et/ou humains, restrictions imposées aux annonceurs récidivistes) empêcheraient les contenus et communications commerciales véhiculant de fausses information d'exister sur leur service.

Or la Commission européenne, sur le fondement de travaux académiques récents, a souligné que malgré des mesures prises par certains opérateurs, des problèmes persistaient en ce domaine[54]. Les plateformes ont un rôle clé pour empêcher les pourvoyeurs de fausses informations de s'enrichir, notamment en identifiant ces derniers et en empêchant de rattacher une annonce publicitaire ou un contenu sponsorisé à leurs contenus. La collaboration et l'échange d'informations entre elles en ce domaine sont à encourager.

5.5.5. Les mesures relatives à la promotion de contenus d'information se rattachant à un débat d'intérêt général

La question de la promotion de contenus d'information se rattachant à un débat d'intérêt général revêt un **enjeu démocratique majeur.** Ces contenus peuvent en effet contribuer à former l'opinion des citoyens. Lutter contre les fausses informations dans ce contexte est ainsi essentiel pour préserver le pluralisme des opinions.

Si l'ensemble des plateformes reconnaît la particularité et l'importance de ces contenus, elles en ont des approches différentes. Il apparaît donc important d'œuvrer à l'harmonisation[55] de la définition de cette notion.

Les mesures prises en ce domaine sont particulièrement variées : certaines plateformes interdisent de manière permanente la promotion de ces contenus ; d'autres appliquent des restrictions de ciblage ; d'autres encore distinguent ces promotions des autres publicités par le biais d'une labellisation particulière comportant parfois davantage d'informations sur

[54] Communication de la Commission européenne du 26 mai 202, COM(2021) 262 final, « Orientations visant à renforcer le code de bonnes pratiques contre la désinformation ».
[55] Cette problématique avait déjà été relevée par le groupe des régulateurs européens des services de médias audiovisuels (ERGA) dans les rapports suivants : *Report of the activities carried out to assist the European Commission in the intermediate monitoring of the Code of practice on disinformation*, juin 2019, et *ERGA report on disinformation: Assessment of the implementation of the code of practice*, mai 2020.

l'annonceur. Ces initiatives répondent à une logique de transparence et d'information des utilisateurs et ne peuvent être qu'encouragées.

De nombreux acteurs ont mis en place un processus d'identification particulier des annonceurs voulant faire la promotion de contenus d'information se rattachant à un débat d'intérêt général. Une labellisation particulière existe sur Facebook et Instagram : les publicités d'enjeu social, électoral et politique doivent être clairement labellisées avec la mention « *Payé par* ». La pertinence de ces mesures, qui peuvent varier d'une plateforme à l'autre[56], est réelle, en particulier dans l'objectif de parer de possibles influences, notamment étrangères, dans le débat démocratique interne.

Bien que des initiatives allant dans le bon sens soient prises, elles semblent être toujours insuffisantes[57]. Le CSA note que **les répondants n'ont pas davantage coopéré entre eux pour atteindre une plus grande harmonisation des approches retenues en matière de promotion de contenus d'informations se rattachant à un débat d'intérêt général.** Il appelle les plateformes à être davantage proactives et à poursuivre leurs efforts, notamment au regard des récentes annonces d'une future législation européenne en matière de publicités politiques et à l'approche de l'élection présidentielle de 2022 en France.

5.5.6. La mise en place de bibliothèques publicitaires

La mise en place, par certaines plateformes (Facebook, Google, Twitter, Snapchat) de bases de **données publiques relatives aux contenus publicitaires** est une avancée importante. Certaines d'entre elles comportent des informations sur les communications commerciales en général et pas seulement sur contenus promu d'information se rattachant à un débat d'intérêt général, ce qui répond à une logique de **transparence nécessaire** et doit, pour cela, être encouragé. En effet, bien que la forme, le contenu et les filtres de recherches varient beaucoup, ces outils sont des éléments clés pour améliorer la transparence mais également pour comprendre l'impact des publicités.

Cependant, les informations contenues dans ces bases de données ne semblent pas toujours suffisantes. Dans deux rapports relatifs à la lutte contre la désinformation[58] publiés en 2019 et en 2020, le groupe des régulateurs européens des services de médias audiovisuels (ERGA) relève que ces informations sont agrégées, parfois difficilement trouvables ou filtrées, et qu'il est parfois impossible d'en garantir la fiabilité. Il est également

[56] Cf. rapports ERGA précités.
[57] Cela avait notamment été relevé par l'ERGA dans un communiqué de presse du 16 avril 2021 : https://erga-online.eu/wp-content/uploads/2021/04/210416_PR_ERGA_PoliticalAdvertising.pdf
[58] *Ibid.*

important que les plateformes qui ne le permettent pas encore rendent accessibles ces bases de données en français et permettent de filtrer les résultats pour la France.

Au vu de ces éléments d'analyse, le CSA formule les préconisations suivantes :

- **Rendre les politiques publicitaires accessibles en français et faire en sorte qu'elles comportent une partie relative à la manipulation de l'information.**
- **Mettre en place des outils faciles d'accès et d'utilisation permettant à l'utilisateur de comprendre pourquoi il a été ciblé dans une situation donnée et de paramétrer ses préférences publicitaires.**
- **Mettre à disposition du Conseil (à tout le moins) des données chiffrées permettant de saisir l'ampleur des liens financiers entre manipulation de l'information et communications commerciales sur les plateformes**..

5.6. Éducation aux médias et à l'information et relations avec le monde de la recherche

Sur les onze déclarants, quatre ont rendu compte d'actions concrètes d'ampleur en matière d'éducation aux médias et à l'information (EMI) sur les thématiques liées à la lutte contre la diffusion de fausses informations (Facebook, Google, Microsoft et Twitter). Trois se sont mobilisés spécifiquement à l'occasion de l'épidémie de Covid-19 (LinkedIn, Webedia (Jeuxvideo.com) et Unify (Doctissimo)) et une a indiqué travailler à la mise en œuvre d'une section dédiée sur son service dans l'année 2021 (Webedia). La Fondation Wikimédia fait valoir que « *la promotion de l'éducation aux médias et à l'information est au cœur même [de son] projet* ». Verizon Media n'a pas abordé cet aspect dans sa déclaration.

5.6.1. Le développement d'initiatives à destination des professionnels de l'information

Cette année encore, les actions menées par les opérateurs en matière d'EMI s'adressent principalement aux jeunes publics, de 6 à 20 ans. Lorsque les adultes sont visés, c'est le plus souvent en tant que parents plutôt que dans le cadre de leurs propres usages de la plateforme. À titre d'exemple, Google a développé en mars 2020, l'application « *Family Link* », qui visait à aider les parents à maîtriser l'utilisation d'Internet de leurs enfants pendant le confinement. L'application leur permettait de déterminer des horaires d'utilisation ou encore d'approuver ou de refuser le téléchargement d'applications sur le téléphone de ces derniers.

Ce constat reste à nuancer : on note par exemple des initiatives interactives amenant les utilisateurs de tout âge à s'interroger pour mieux savoir repérer les fausses informations, telle que la campagne « *Trois questions pour éradiquer les* fake news » de Facebook[59].

La nouveauté relevée en 2020 concerne la **multiplication d'initiatives à destination des journalistes.** En effet, avec l'épidémie de Covid-19 et la tenue des élections présidentielles aux États-Unis et municipales en France, certaines plateformes ont proposé des formations ou pris des mesures pour faciliter le travail des rédactions. Google s'est particulièrement distingué avec la *Google News Lab* qui a organisé des ateliers de formation à la vérification dans les États membres de l'Union Européenne. Ainsi, en amont du premier tour des élections municipales en France, les équipe de Google se sont déplacées dans les rédactions ou ont reçu des journalistes dans des ateliers en région.

Les actions des opérateurs en matière d'EMI sont très souvent développées en partenariats avec des tiers (associations, agences publiques, centres de recherches…), dynamique favorable qui permet d'apporter l'expertise du partenaire et de lui donner une forte visibilité.

5.6.2. Des données sur les audiences touchées mais toujours très peu d'informations sur l'impact et le coût de ces dernières

Les opérateurs ont été plus nombreux qu'en 2019 à communiquer dans leur déclaration des données concernant le nombre de personnes touchées par leurs actions : Facebook, Google, Webedia.

En revanche, une seule plateforme, Facebook, a transmis des informations rendant compte de l'impact d'une de ses initiatives sur le comportement des utilisateurs, la campagne « *Trois questions pour éradiquer les* fake news ». En effet, l'opérateur a indiqué avoir mené une étude pour mesurer l'efficacité de son action auprès de « groupes test » et de « groupes de contrôle », ce qui a permis de mettre en avant une meilleure mémorisation des annonces diffusées sur son service et un changement de comportement après avoir interagi avec la campagne.

Pour la deuxième année consécutive, la grande majorité des plateformes n'a communiqué aucune information sur le coût des actions et les moyens qui leur sont consacrés. Seuls Facebook et Google donnent des éléments de budget, notamment, le premier, au sujet du Fonds pour le civisme en ligne qui a financé 20 initiatives en 2020 en France à hauteur

[59] Microsoft a également fait état de mesures de ce type (quiz interactifs sur la vie démocratique dans le cadre de la campagne présidentielle américaine ou sur les *deepfakes*) mais qui visent le public américain.

d'un million d'euros et le second, au sujet de bourses d'un million d'euros octroyées à plusieurs initiatives de formation en France[60].

5.6.3. Des rapprochements notables mais encore perfectibles avec le secteur de la recherche s'agissant de l'exploitation de la data

Cinq opérateurs ont déclaré avoir mené des actions avec le monde de la recherche, de façon bien plus détaillée et précise que l'année passée : Facebook, Microsoft, Twitter, la Fondation Wikimédia et Google.

Facebook mène un certain nombre d'actions, parmi lesquelles on peut citer le lancement d'une chaire universitaire d'EMI en partenariat avec l'École supérieure de journalisme de Lille, le développement de la plateforme FORT qui fournit aux universitaires et chercheurs des outils et données utiles pour étudier l'impact de la plateforme sur la démocratie, les élections et le bien-être, ou encore un partenariat de recherche pour étudier l'impact de Facebook et Instagram sur les comportements politiques dans le cadre de l'élection présidentielle américaine. Par ailleurs, l'opérateur précise qu'il participera, en 2021, à un groupe de travail mis en place par l'Observatoire européen des médias numériques (European Digital Media Observatory – EDMO) qui vise à élaborer un code de conduite, conformément à l'article 40 du règlement général sur la protection des données (RGPD), afin de faciliter le partage responsable des données, y compris celles sur les plateformes numériques, à des fins de recherche en sciences sociales.

Twitter a procédé à une mise à jour de son API afin notamment de permettre aux chercheurs d'obtenir un accès gratuit à l'historique complet des conversations publiques, des niveaux d'accès plus élevés et gratuits à la plateforme de développement Twitter, des capacités de filtrage plus précises ou encore de nouveaux guides techniques et méthodologiques.

Le CSA constate l'investissement de Google dans la collaboration avec les chercheurs sur l'étude de la diffusion et l'impact de la manipulation de l'information, notamment par la facilitation de l'accès aux informations sur les algorithmes de classement de recherche et la mise à disposition de données. Par ailleurs, la plateforme annonce également la création d'un projet sur l'EMI des citoyens dans la lutte contre la désinformation en Europe par l'investissement dans le lancement du fonds européen pour les médias et l'information sur les cinq années à venir.

[60] Une autre plateforme fournit le budget d'une action, mais le caractère confidentiel de l'information rend son intérêt moindre dans le cadre du présent bilan.

Microsoft mène de nombreuses initiatives avec le monde de la recherche, aussi bien en sciences humaines que sur le développement d'outils pour la lutte contre la manipulation de l'information.

Par ailleurs, il a également été ajouté cette année au sein du questionnaire un focus sur les *deepfakes* et les moyens de lutte contre les trucages vidéo à des fins de propagation de fausses informations. Le CSA constate des initiatives de Facebook, Microsoft et Google en la matière aussi bien dans la contribution aux jeux de données en open source permettant d'entraîner les algorithmes de détection que dans la collaboration avec les chercheurs pour développer des outils**. Le Conseil salue la prise en compte des enjeux de ces technologies par les plateformes ainsi que la mise en place de projets multidisciplinaires impliquant conjointement plusieurs opérateurs**, notamment le partenariat sur l'intelligence artificielle réunissant Microsoft, Facebook, Google et la Fondation Wikimédia.

Au vu de ces éléments d'analyse, le CSA formule les préconisations suivantes :

- **Évaluer l'impact sur le comportement des utilisateurs des actions en matière d'éducation aux médias et à l'information** et en rendre compte au public et au régulateur.
- **Intensifier les collaborations avec le monde de la recherche, notamment en mettant en œuvre les conditions permettant la mise à disposition et l'exploitation des données à grande échelle.**

Conclusion

Le présent bilan atteste des efforts fournis par les opérateurs par rapport à l'exercice 2019, notamment pour faire face à l'abondance des phénomènes de désinformation liés à la crise sanitaire. Il tient en outre à saluer l'esprit de coopération et la disponibilité dont a fait preuve la majorité des opérateurs, bien qu'il constate des lacunes notables et une grande hétérogénéité dans les informations déclarées, d'un opérateur à un autre, et selon les sujets traités. Il demande aux opérateurs d'œuvrer à davantage de transparence dans la nature, la précision et la clarté des informations fournies au public et au régulateur.

Le **dispositif de signalement** a été mis en place sur l'ensemble des plateformes, à l'exception de Wikipédia[61], avec une accessibilité, une visibilité et une clarté qui demeurent néanmoins inégales. Si les opérateurs mettent en avant leur volonté d'améliorer l'expérience utilisateur de ce dispositif, le Conseil constate que ce dernier manque parfois de clarté et de visibilité, notamment dans les moteurs de recherche.

Concernant **la transparence des algorithmes**, le CSA remarque une hausse notable de la quantité d'informations fournies par certains opérateurs, sans pour autant que ces informations soient suffisantes pour permettre une évaluation de l'effectivité des systèmes algorithmiques de recommandation et de modération dans la lutte contre la manipulation de l'information. Le développement de quelques fonctionnalités permettant la transmission d'informations aux utilisateurs, notamment de manière contextuelle et personnalisée, sur le fonctionnement et les effets de ces systèmes est à saluer ; néanmoins, ce mécanisme reste d'une manière générale très insuffisant, voire inexistant sur certains services.

Les nouvelles mesures déclarées en matière de **promotion des contenus issus d'entreprises et d'agences de presse et de services de communication audiovisuelle** s'inscrivent directement dans le contexte de la crise sanitaire et portent principalement sur l'identification des sources. La nécessité de faire face aux phénomènes de désinformation et de mésinformation autour de la Covid-19 a amené les opérateurs à engager des collaborations avec les pouvoirs publics, des experts en matière de désinformation, la communauté scientifique et les organisations nationales et internationales. Ces collaborations portent notamment sur la fourniture d'informations de contexte et promouvoir des contenus issus de source d'autorité.

De la même manière, de nouvelles initiatives ont été prises par certains opérateurs à l'encontre des différentes **pratiques de propagation massive de fausses informations** et

[61] Ce qui s'explique par le mode de fonctionnement coopératif de son service, basé sur l'actualisation constante des contenus par les contributeurs et la modération participative.

dans l'étude des opérations d'influence coordonnées. Néanmoins, le CSA remarque le manque d'informations communiquées aux utilisateurs sur les risques qui en découlent et préconise d'intensifier le travail de collaboration entre les acteurs pour lutter contre ce type de pratiques.

Le Conseil constate une légère amélioration des moyens mis en œuvre **dans la lutte contre les communications commerciales porteuses de fausses informations et en matière de promotion des contenus d'information se rattachant à un débat d'intérêt général**, notamment dans la mise en place de bibliothèques publicitaires. Il en appelle cependant à la communication de davantage de données chiffrées afin de mieux appréhender les risques en présence.

Enfin, davantage d'initiatives **d'éducation aux médias et à l'information** ont été déclarées, sans que leur impact réel ne soit renseigné. Certains opérateurs ont développé les collaborations avec **le monde académique**, notamment dans la lutte contre la désinformation et sur l'ouverture des données aux chercheurs. Sur ce dernier sujet, le Conseil salue également la collaboration entre les opérateurs.

Certaines nouvelles mesures vont dans le sens de préconisations formulées par le Conseil dans son bilan de l'année passée. Le Conseil en formule de nouvelles dont il s'attachera, dans son prochain rapport, à évaluer la prise en compte par les opérateurs. De manière générale, la transparence à l'égard du régulateur et du public est une des clés de voûte d'un dispositif de lutte contre les contenus et comportements préjudiciables ou illégaux. Elle est d'ailleurs fondamentale dans le projet de *Digital Services Act*, en discussion au sein de l'Union européenne. Le Conseil appelle les opérateurs à prendre toute la mesure de cet objectif en fournissant une information claire, substantielle et adaptée aux besoins des différents publics.

Les mois qui suivront la publication du présent bilan seront marqués par un contexte particulier, outre la crise sanitaire en cours : la France s'apprête à entrer dans deux périodes électorales jusqu'aux scrutins présidentiels et législatifs de 2022. Cette période sera propice au développement de phénomènes de désinformation susceptibles de nuire à la sincérité des scrutins. Aussi, le Conseil portera une attention toute particulière aux mesures déployées par les opérateurs pour prévenir et, le cas échéant, contrer ces risques, comme à la préservation de la liberté de communication.

Synthèse des préconisations du Conseil

Sur la transparence en général

- **Préconisation n° 1 : fournir de manière proactive aux utilisateurs, sur la plateforme, si possible de manière personnalisée et contextuelle, des explications claires et accessibles** sur les mesures mises en œuvre face aux risques liés à la manipulation de l'information.

- **Préconisation n° 2 : faire preuve de plus de transparence vis-à-vis du public en fournissant davantage de précisions chiffrées et d'éléments contextualisés, notamment dans les déclarations**, et communiquer au Conseil toutes les informations, fussent-elles confidentielles, permettant de mieux comprendre les mesures prises et leur impact.

Sur le dispositif de signalement

- **Préconisation n° 3 :** pour les moteurs de recherche (Google Search, Yahoo Search et Bing), **améliorer la visibilité et la facilité d'utilisation de leur dispositif de signalement.**

- **Préconisation n° 4 :** mieux informer les auteurs de signalement, et utilisateurs ayant publié un contenu signalé, de **l'avancée des procédures de traitement des signalements en cours, et leur en communiquer l'issue dans un délai raisonnable.** Par ailleurs, il apparaît impératif de **mieux expliquer aux utilisateurs les voies de recours existantes.**

- **Préconisation n° 5 :** à l'exception de la lutte contre certaines pratiques aisément détectables automatiquement, **maintenir une intervention humaine dans le processus de décision d'une action à l'égard d'un contenu ou d'un compte**.

Sur la transparence des algorithmes

- **Préconisation n° 6 : proposer aux utilisateurs des fonctionnalités leur permettant de comprendre**, si possible de manière personnalisée et contextuelle, **les effets des systèmes algorithmiques de recommandation et de modération**.

- **Préconisation n° 7 : déclarer tout élément donnant à voir comment la lutte contre les biais est concrètement mise en œuvre sur leurs services** : ressources dédiées, outils, modifications opérées par la suite, résultats.

Sur la promotion des contenus issus d'entreprises et d'agences de presse et de services de communication audiovisuelle

- **Préconisation n° 8** : **conserver une part d'intervention humaine dans la vérification des informations fournies au sujet des organes de presse lorsqu'ils font l'objet d'une identification particulière**.

- **Préconisation n° 9 : développer des initiatives et partenariats tels que ceux pris dans le contexte de la crise sanitaire** (en coopération avec des pouvoirs publics, des associations, des chercheurs, etc.) tout en veillant, le cas échéant, à distinguer les sources gouvernementales, à des fins de transparence vis-à-vis des utilisateurs.

Sur la lutte contre les comptes propageant massivement de fausses informations

- **Préconisation n° 10 : informer davantage les utilisateurs sur les pratiques coordonnées d'influence et les risques qui en découlent,** tout particulièrement en période électorale.

- **Préconisation n° 11 : faire preuve de davantage de transparence, à tout le moins à l'égard du régulateur, sur les recettes publicitaires potentielles générées par les comptes propageant massivement de fausses informations.**

Sur les mesures de lutte contre les fausses informations en matière de communications commerciales et de promotion de contenus d'information se rattachant à un débat d'intérêt général

- **Préconisation n° 12 : rendre les politiques publicitaires accessibles en français et faire en sorte qu'elles comportent une partie relative à la manipulation de l'information.**

- **Préconisation n° 13 : mettre en place des outils faciles d'accès et d'utilisation permettant à l'utilisateur de comprendre pourquoi il a été ciblé dans une situation donnée et de paramétrer ses préférences publicitaires**.

- **Préconisation n° 14 : mettre à disposition du Conseil (à tout le moins) des données chiffrées permettant de saisir l'ampleur des liens financiers entre manipulation de l'information et communications commerciales sur les plateformes**.

Sur l'éducation aux médias et à l'information ainsi que sur les relations avec le monde de la recherche

- **Préconisation n° 15** : **évaluer l'impact sur le comportement des utilisateurs des actions en matière d'éducation aux médias et à l'information** et en rendre compte au public et au régulateur.

- **Préconisation n° 16 : intensifier les collaborations avec le monde de la recherche, notamment en mettant en œuvre les conditions permettant la mise à disposition et l'exploitation des données à grande échelle.**

Annexe : typologie des mesures prises par les plateformes pour lutter contre la manipulation de l'information

1. Dispositif de signalement

DISPOSITIF DE SIGNALEMENT	
Forme, accessibilité, visibilité	**Outil à proximité immédiate du contenu** : Dailymotion, YouTube, Facebook/Instagram, LinkedIn, Twitter, Jeuxvideo.com, Snapchat, Doctissimo ➔ **accessible uniquement aux utilisateurs connectés :** Facebook/Instagram, LinkedIn, Twitter, Jeuxvideo.com, Snapchat ➔ **signalé par un symbole :** Facebook/Instagram, YouTube, LinkedIn, Twitter, Jeuxvideo.com, Snapchat, Doctissimo. ➔ **signalé par un mot** avec un hyperlien (« Signaler ») : Dailymotion. **Existence d'un formulaire de signalement** depuis le bas de la page ou depuis le site institutionnel : Google (formulaire web standard pour le signalement), Microsoft (Bing), Yahoo, Snapchat, Doctissimo ➔ **accessible uniquement aux utilisateurs connectés :** Facebook **Outil de signalement proposé** : ➔ **pour tous les contenus (publications, publicités, chaînes, comptes, commentaires) :** Dailymotion, Facebook/Instagram, LinkedIn, Twitter, Snapchat ; ➔ **pour certains contenus** : YouTube, Jeuxvideo.com, Doctissimo, Yahoo. **Outil accessible en** : ➔ **un clic** : Dailymotion (3 étapes), YouTube, Facebook/Instagram (3 étapes), LinkedIn (3 étapes), Jeuxvideo.com pour les utilisateurs connectés, Snapchat, Twitter, Doctissimo ; ➔ **deux** clics : Jeuxvideo.com (pour les utilisateurs non connectés) **Dispositif de signalement via le centre d'aide et le formulaire de contact** : Google, Microsoft (Bing) YouTube
Motifs proposés	**Motif :** ➔ **« Fausse information** » ou formulation équivalente : Facebook/ Instragram, Snapchat, Doctissimo ; ➔ **« Désinformation »** : Dailymotion, LinkedIn ; ➔ **« Fausse information au sujet d'élections »** ou formulation équivalente : Twitter, LinkedIn ; ➔ **« Spam ou contenu trompeur »** : YouTube **Absence de motif « fausse information » ou s'y rattachant :** Google (moteur de recherche), Microsoft (Bing), Jeuxvideo.com
Forme de l'arborescence	**Arborescence** : ➔ tous les motifs sont visibles en même temps : Dailymotion, YouTube, Microsoft Advertising, Jeuxvideo.com, Doctissimo ; ➔ par strates : Facebook, LinkedIn, Twitter, Snapchat.

	Apparition du motif « fausse information » (ou assimilé) : ➔ **au niveau 1 de l'arborescence** : Dailymotion ; YouTube, Facebook, Snapchat, Doctissimo ; ➔ **au niveau 2 de l'arborescence :** LinkedIn, Snapchat ; ➔ **absent avec champ « autre »** : Bing, Jeuxvideo.com (utilisateurs non connectés) ; Google ; ➔ **absent sans champ « autre » :** Twitter (quand le contenu induit en erreur au sujet d'élections), Jeuxvideo.com (utilisateurs connectés).
Fonctionnalités de l'outil	**Choix parmi** : ➔ **items** uniquement : Facebook, YouTube (application), Bing, LinkedIn, Snapchat ; ➔ items et **zone de texte** libre : Dailymotion, YouTube (version ordinateur), Microsoft Advertising, Jeuxvideo.com, Doctissimo ➔ items, zone de texte libre et possibilité d'**ajout d'informations complémentaires** (ex. : captures d'écran) : Twitter. Possibilité de **signaler plusieurs contenus** en même temps : YouTube, Twitter
Nombre de signalements pour fausses informations	**Chiffres déclarés** : - 0 pour Google ; - 473 en France pour Dailymotion ; - 4 789 pour Doctissimo ; - 24 919 pour LinkedIn (en France) ; - 50 000 environ pour Jeuxvideo.com ; - 168 709 via le canal « *fausses informations susceptibles d'altérer la sincérité d'un scrutin ou de troubler l'ordre public* » (France) pour Twitter. **Absence de données déclarées** : Facebook, Bing. **Données déclarées de manière confidentielle** : Snapchat

SANCTION DES FAUSSES INFORMATIONS	
Politique de traitement des fausses informations	**Externalisation** du traitement par des **tiers vérificateurs de confiance** (*fact-checkers*) : Facebook, Twitter (en période électorale). **Traitement par une équipe de modérateurs :** ➔ **dédiée :** Dailymotion; ➔ **non dédiée** : Google, LinkedIn, Jeuxvideo.com, Snapchat, Doctissimo. **Approche spécifique :** ➔ Wikipédia : **modération par les utilisateurs ;** ➔ Twitter : **modération uniquement des fausses informations en matière électorale.**
Normes de référence du contrôle des plateformes	**Loi locale**[62] **:** Microsoft, Jeuxvideo.com. **Règles et standards de la communauté**, adaptés pour tenir compte de la loi : Dailymotion, Facebook, LinkedIn, Twitter, Snapchat, Jeuxvideo.com.

[62] Loi du 22 décembre 2018, en l'espèce.

Procédure d'instruction des signalements	**Ouverture et traitement d'un signalement** : ➔ **traitement de tous les signalements :** Dailymotion, Jeuxvideo.com, Snapchat, Doctissimo. ➔ traitement des signalements **en fonction de critères**: LinkedIn (précisés dans la déclaration), Facebook (non précisés) ; ➔ traitement **uniquement des signalements de fausses informations au sujet d'élections** : Twitter ; ➔ **pas d'indication** : Google, Microsoft. **<u>Procédures de priorisation</u>** : en période électorale ou d'urgence (Dailymotion), avec le programme *Trusted Flaggers* (traitement prioritaire des signalements émanant d'acteurs identifiés au regard de leurs compétences) (YouTube).
Types de mesures adoptées	**Mesures sur le contenu** : ➔ **retrait :** Dailymotion, Google, Facebook/Instagram (en période électorale Doctissimo, Microsoft Advertising et Bing (quand contenu en violation de la législation locale), Twitter (contenu faux et nuisible au sujet des vaccins contre la Covid-19 ou destiné à saper la confiance du public dans une élection ou tout autre processus civique) ; ➔ **désindexation** de l'URL lorsque la loi l'exige : Bing ; ➔ **réduction de la visibilité** : Google, Facebook ➔ **information de l'utilisateur** sur le fait qu'il est en train d'interagir avec un contenu diffusant de fausses informations : YouTube, Facebook/Instagram, Twitter (informations trompeuses sur la Covid-19, ou destinées à saper la confiance du public dans une élection ou tout autre processus civique, ou incluant des *deepfakes*). ➔ **Indiqué de manière confidentielle :** Facebook, LinkedIn **Mesures sur le compte :** ➔ **suppression du compte** : Facebook, Doctissimo, Jeuxvidéo.com, LinkedIn, Twitter ; ➔ **mesures disciplinaires à l'encontre du compte** : LinkedIn, Twitter, Jeuxvideo.com, Doctissimo (suspension temporaire ou définitive), Wikipédia (blocage en écriture).
Nombre d'actions sur les comptes et contenus	**Globalement peu d'informations sur le nombre de contenus retirés** après avoir été signalés comme propageant une fausse information. **-** Dailymotion : 5 comptes clôturés pour propagation de fausses informations.
Délai d'instruction des signalements	Délai de traitement moyen : Dailymotion (11h) Autres opérateurs : pas d'information sauf Jeuxvideo.com de manière confidentielle.

INFORMATION DES UTILISATEURS	
Information de l'ensemble des utilisateurs	Informations générales délivrées sur : - la **procédure** pour effectuer un signalement (ex : Facebook) ; - les **règles de fonctionnement** (autorisations/interdictions) ; - les **sanctions** encourues ; - les **contenus visés par une décision** de l'opérateur (contenu masqué, contenu désindexé, contenu retiré ; ex. : YouTube, LinkedIn).
Information après un signalement	**Informations après un signalement délivrées à** : ➔ **l'auteur du signalement** : suivi de la demande, instruction, décision, recours : Facebook, LinkedIn, Twitter, Jeuxvideo.com, Doctissimo ; ➔ **l'auteur du contenu signalé :** notification du signalement, décision, recours : LinkedIn, Twitter, Jeuxvideo.com ; ➔ **tous les utilisateurs** : Doctissimo (sanction envers le compte ou le contenu indiquée à l'emplacement du message initial). **Informations après le retrait du contenu signalé :** Dailymotion. **Pas de dispositif** d'information : Bing, Snapchat.
Existence de voies de recours	**Pour :** ➔ **l'auteur du contenu** : Dailymotion (uniquement aux partenaires professionnels), Facebook, LinkedIn, Twitter, Jeuxvideo.com, Doctissimo (par mail) ; ➔ **l'auteur du signalement :** Facebook, LinkedIn, Jeuxvideo.com.
L'ouverture de l'instruction d'un signalement ne semble pas faire l'objet d'une information particulière auprès des utilisateurs.	

2. Transparence des algorithmes

SYSTÈMES DE RECOMMANDATION DE CONTENUS	
Informations communiquées au régulateur	
La ou les **fonctionnalités** du service auquel s'appliquent les systèmes sont clairement énoncées	Dailymotion, Facebook, LinkedIn (fil d'actualité), Google (requêtes sur Search et YouTube, page d'accueil et « A suivre » sur YouTube), Snapchat (Découvrir, Stories et Spotlight), Twitter (fil d'actualité, section « Au cas où vous l'auriez manqué », classement des Tweets dans une conversation, suggestion de comptes), Wikipédia (moteur de recherche interne), Yahoo (résultats de recherche). Bing estime ne pas faire de recommandation de contenus.
L'**objectif principal** de ces systèmes est porté à la connaissance du régulateur (tâche ou classe de problèmes à résoudre)	Il s'agit notamment: - d'indexer du contenu : Dailymotion, Snapchat ; - de classer du contenu : Dailymotion, Facebook, LinkedIn (fil d'actualité), Google (Search, YouTube), Snapchat (Découvrir), Yahoo ; - d'afficher du contenu selon sa pertinence : Twitter, Wikipédia.
Le ou les **types d'algorithmes** utilisé(s) par ces systèmes sont communiqués	Apprentissage automatique : Google (Search et YouTube) Traitement automatique du langage naturel : Google, Wikipédia. Pas d'information : Facebook, LinkedIn, Snapchat, Twitter.
Les **données prises en entrée** ainsi que les **données de sortie** de ces systèmes sont renseignées	Oui : Wikipédia (algorithme de recherche interne). Oui mais non-exhaustives : Facebook, Twitter, LinkedIn (fil d'actualité), Google, Yahoo (algorithmes de recherche), Snapchat.
Les éventuelles modalités de l'**intervention humaine** dans ces systèmes sont bien identifiées	Tests réalisés auprès d'utilisateurs ou d'évaluateurs externes : Google (Search et YouTube)
Traitement (spécifique ou indifférencié) des **fausses informations** dans ces systèmes	Oui : Dailymotion (désindexées le temps de l'instruction)
Informations communiquées aux utilisateurs	
L'ensemble des **informations précédentes** est communiqué aux utilisateurs	Oui : LinkedIn, Twitter (langage simple et détaillé dans le Centre d'aide), Snapchat (centre de confidentialité) Oui, par type de public : Dailymotion (développeurs, utilisateurs, visiteurs), Google (information spécifique aux propriétaires de sites).
Information des utilisateurs de l'existence de **réglages permettant**	Oui, via des informations proactives : Snapchat Oui, à proximité des contenus : Facebook (publicitaires

de personnaliser la manière dont les contenus leur sont recommandés	et organiques), Snapchat (Découvrir, Stories). Oui, dans un espace d'aide : Facebook (données collectées hors du service), Google (paramètres du compte Google), Microsoft (paramètres de confidentialité pour Advertising), Snapchat (réglages de l'application, centre de confidentialité), Twitter (sans précision). Non (pas de paramètres de réglage spécifiques) : Dailymotion, LinkedIn, Wikipédia. Pas d'information : Yahoo.
Information aux utilisateurs des mises à jour des systèmes affectant significativement la manière dont les contenus sont recommandés	Oui, de manière proactive : Facebook (pages d'aide et actualités de l'entreprise), Google (sur Search pour les propriétaires de site web).
Possibilité pour les utilisateurs de **demander des informations supplémentaires** sur les systèmes de recommandation utilisés	Aucune information communiquée par les opérateurs à ce sujet
Explication locale sur les résultats des systèmes de recommandation	Oui : Facebook (« Activité sur Facebook », publicités et contenus organiques), Microsoft (Advertising).

SYSTÈMES DE MODÉRATION DE CONTENUS	
Informations communiquées au régulateur	
La ou les **fonctionnalités** du service auquel s'appliquent les systèmes sont clairement énoncées	Dailymotion (algorithmes de priorisation des signalements pour fausses informations « *en particulier durant les périodes électorales ou d'urgence* », algorithme d'indexation qui désindexe les contenus en cours de modération), Twitter (Tweets cachés si considérés comme abusifs), Wikipédia (évaluation des contributions et des contributeurs), Facebook (identification de contenus avant intervention).
L'**objectif principal** de ces systèmes est porté à la connaissance du régulateur (tâche ou classe de problèmes à résoudre)	Evaluer les contributions ainsi que les contributeurs pour lutter contre le vandalisme : Wikipédia. Analyse et traitement du langage naturel : Facebook, Twitter, Wikipédia (ORES).
Le ou les **types d'algorithmes** utilisé(s) par ces systèmes sont communiqués	Outils automatiques de suppression de certains contenus : Snapchat, Twitter. Intelligence artificielle (sans précision sur la nature) : Facebook (SimsSearchNet++, ObjectDNA, LASER). Apprentissage automatique (*machine learning*) : Dailymotion (algorithmes d'indexation), Facebook (sans plus de précision), Twitter (détection proactive, sans plus de précision), Wikipédia (ClueBot NG). Apprentissage profond (*deep learning*) : Facebook (détection des *deepfakes* basée sur EfficientNet). Réseaux antagonistes génératifs (GAN) : Facebook (détection de *deepfakes*).

Les **données prises en entrée** ainsi que les **données de sortie** de ces systèmes sont renseignées	Oui, pour partie : Facebook (réutilisation des évaluations des partenaires médias pour améliorer les systèmes).
Les éventuelles modalités de l'**intervention humaine** dans ces systèmes sont bien identifiées	Recours à des évaluateurs externes : YouTube. Recours à des experts certifiés (ex. médecins) : YouTube. Contenus examinés par des équipes dédiées : Twitter, Snapchat
Traitement (spécifique ou indifférencié) des **fausses informations** dans ces systèmes	Dailymotion (priorisation des signalements)
La **performance** des systèmes de modération est communiquée	Oui : Wikipédia (ClueBot NG : 90 % de contributions classées correctement).
Informations communiquées aux utilisateurs	
L'ensemble des **informations précédentes** est communiqué aux utilisateurs	Oui, dans des pages dédiées : Facebook (blog Engineering), Snapchat (règles communautaires), Twitter (conditions d'utilisation, centre de transparence).
Informations des utilisateurs de l'existence de **réglages permettant de personnaliser** la manière dont les contenus sont modérés	Oui : Twitter (gestion des notifications reçues)
Information aux utilisateurs des mises à jour des systèmes affectant significativement la manière dont les contenus sont modérés	Aucune information communiquée par les opérateurs à ce sujet
Possibilité pour les utilisateurs de **demander des informations supplémentaires** sur les systèmes de modération utilisés	Aucune information communiquée par les opérateurs à ce sujet
Explication locale sur les décisions de modération	Aucune information communiquée par les opérateurs

SYSTEMES SPECIFIQUES A LA LUTTE CONTRE LA MANIPULATION DE L'INFORMATION	
Utilisation de systèmes algorithmiques dans la détection et le traitement de fausses informations	Dailymotion (signalements priorisés), Snapchat (outil de détection des termes abusifs dont termes liés à la COVID-19), Twitter (identification avant intervention), Facebook, LinkedIn (démonétisation ou masquage du contenu problématique).
Lutte contre les campagnes coordonnées de désinformation	Facebook (comportements inauthentiques coordonnés), Twitter (opérations de désinformation), Wikipédia (détection de « faux-nez »).

Détection d'hypertrucages ou *deepfakes*	Facebook (réseaux antagonistes génératifs entraînant des algorithmes de détection), Microsoft (Microsoft Video Authenticator, détection de *deepfakes* dans Microsoft Azure, développement d'une norme commune d'authentification avec d'autres acteurs).
Mise en avant de sources fiables	Service tout ou partie fermé aux contenus amateurs : Dailymotion, Snapchat (Découvrir). Remontée des résultats issus de sources de référence : Bing (pour certaines requêtes), Google et YouTube (bandeaux d'information liés à certaines requêtes), LinkedIn (redirection de certaines requêtes liées à la COVID-19 vers des pages LinkedIn dédiées). Informations réputées fiables au sein d'onglets ou parties dédiées du service : Bing (onglets dédiés, *« centres d'information COVID-19 »* sur Microsoft News), Facebook, LinkedIn (pages LinkedIn dédiées alimentées par des sources d'autorité), YouTube (sections « A la une », « Actualités » et « A suivre »).
Rétrogradation de fausses informations	Dailymotion (désindexation des contenus signalés pour fausses informations le temps de l'instruction ; les fausses informations repérées ne sont plus référencées ni recommandées[63]), Bing, Facebook.
Renforcement des contrôles préalables à l'utilisation de certains services	Bing (interdiction de certaines campagnes publicitaires liées à la COVID-19, limitations de l'utilisation de Custom Neural Voice[64] pour *« contrer la prolifération des deepfakes »*).

[63] « *Les fausses informations repérées (…) recommandées* » : information confidentielle.
[64] Outil d'intelligence artificielle conçu par Microsoft permettant de créer une voix artificielle.

3. Promotion des contenus issus d'entreprises et d'agences de presse et de services de communication audiovisuelle

IDENTIFICATION DES SOURCES DES CONTENUS	
Mesures relatives à l'identification des sources	**Identification réalisée en interne** : o examen humain à l'aune de critères préétablis conduisant à la certification du compte ou de la chaine (non réservé aux EAP-SCA) (Dailymotion) ; o sélection et contractualisation des éditeurs de contenus avec la plateforme (Snapchat et Yahoo). **Identification réalisée en externe par les EAP-SCA ou un tiers** : o la source renseigne les informations sur son identité (Twitter, Facebook) ; o les EAP-SCA peuvent « marquer » leurs contenus : fonctionnalité ClaimReview (Google, Microsoft, Facebook, Bing) où l'éditeur peut apposer le libellé « *fact-checking* » à son contenu. **Identification de la source est réalisée par un tiers** : o module NewsGuard (Microsoft) notant la fiabilité des sites d'information à l'aune de neuf critères ; o identification de médias contrôlés par un État (Facebook, Twitter).
Indicateurs visuels	**Apposition d'un visuel (badge, logo, icône, bouton d'information)** : o bouton d'information sur les éditeurs relayant du contenu externe (Facebook, Google Search, YouTube ; Twitter à certaines occasions). o visuel commun à tout type d'utilisateurs (EPA-SCA ou non) : badge certifiant le compte de l'utilisateur comme étant authentique (Twitter) et/ou digne de confiance (Dailymotion). o visuel spécifique aux EAP-SCA : - fonctionnalités permettant d'identifier les contenus provenant d'éditeurs de presse (boutons de contexte permettant d'avoir de plus amples informations sur l'éditeur de Facebook, Instagram, YouTube, Google, Microsoft) ; - identification des contenus de *fact-checking* (Google, Bing). **Identification spécifique des contenus :** o de médias affiliés ou contrôlés par un État (Facebook, Twitter) ; o comportant un média modifié (Twitter).
PRISE EN COMPTE DES DÉMARCHES DE LABELLISATION ET GARANTIE DE LA FIABILITÉ DES INFORMATIONS	
Sélection de l'EAP-SCA sur le critère de sa participation à une démarche de labellisation	Labellisation conditionnée à : o l'adhésion à une charte éthique ou déontologique ; o la participation à des initiatives de *fact-checking* pour labelliser le compte ou le contenu (Google, Microsoft via Claim Review).

	Possibilité d'intervenir en tant que tiers vérificateur conditionnée à l'appartenance à des organisations reconnues (IFCN49) (Facebook). Mise en avant des seules sources de confiance dans les recommandations aux utilisateurs (Dailymotion)
Règles posées par la plateforme	Lignes directrices que les partenaires sont tenus de suivre (Snapchat). Activité d'édition de contenus en propre (Doctissimo, LinkedIn).
Partenariats et démarches	Participation, notamment financière, à des démarches, initiatives et programmes de recherches pour développer la qualité du travail journalistique numérique : Google (IFCN, GNI, Trust Project…), Twitter (*Disinfo Lab*), Microsoft (*Coalition for Content Provenance and Authenticity*), Snapchat (*Trusted Flaggers*). Mise en place de « hubs médiatiques » rassemblant les informations provenant de sources officielles (Facebook, Snapchat, Bing Google, Doctissimo Yahoo ; Twitter dans des contextes particuliers) Création d'un comité d'experts face aux fausses informations liées à la crise sanitaire : Doctissimo.
MISE EN AVANT DES CONTENUS ISSUS DES EAP-SCA ET NOTAMMENT DE FACT-CHECKING	
Mise en avant des contenus	**Canaux de diffusion dédiés** : page d'accueil ou pages dédiées aux contenus provenant de sources dites « fiables », dont EAP-SCA (Dailymotion, Snapchat, LinkedIn, YouTube, Doctissimo, Bing, Yahoo). **Sélection algorithmique de contenus** provenant de sources dites « fiables » (Google, LinkedIn, Bing, Snapchat, Dailymotion) afin de les faire apparaître de manière préférentielle dans les résultats ou sur les canaux dédiés.
Mise en avant des contenus de *fact-checking*	Invitation à accéder aux contenus des *fact-checkers* lorsqu'on s'apprête à partager un contenu jugé faux par ces derniers (Facebook) Filtres apposés aux contenus jugés faux par les *fact-checkers* (Facebook). Possibilité de rechercher uniquement des contenus *fact-checkés* (Bing, Google).
Fact-checking confié à des tiers vérificateurs	**De manière directe**, avec accès privilégié à la plateforme via des outils dédiés : système faisant remonter les fausses informations signalées aux tiers vérificateurs partenaires ; ceux-ci peuvent publier leurs contenus de vérification en regard du contenu vérifié et jugé comme faux (Facebook).

	<u>Avec un intermédiaire entre les EAP-SCA et la plateforme</u> : NewsGuard signale la fiabilité d'une source par un symbole (bouclier coloré) (Bing). **<u>Marquage par les EAP-SCA de leurs contenus sans accès au « back-office »</u>** : ClaimReview (Google, YouTube, Bing, Facebook).
Fact-checking réalisé en interne	Vérification des informations par une équipe interne s'appuyant sur un guide pratique rédigé par la plateforme ou sur les informations produites par les EAP-SCA faisant autorité (LinkedIn).
Actions sur les contenus de sources réputées peu fiables	Actions contenu par contenu, au cas par cas : • dégradation de la visibilité des contenus jugés faux par des tiers vérificateurs (Google, Bing, Facebook) ; • informations apposées sur le contenu vérifié (Facebook, Instagram). **Pas de restriction technique du partage** des contenus ayant été jugés comme contenant une fausse information.

4. Lutte contre les comptes propageant massivement de fausses informations

DÉFINITION DES PRATIQUES DE DIFFUSION MASSIVE DE FAUSSES INFORMATIONS	
Comptes désignés comme « faux comptes »	-> comptes visant à créer et/ou partager des informations trompeuses (Facebook, Twitter et LinkedIn) -> notamment dans le cadre de « *campagnes d'informations coordonnées* » (*Facebook*) ou dans le but de réaliser des profits (Twitter)
Comptes contrevenant aux règles de la plateforme s'agissant de pratiques trompeuses	**Catégories de comptes dont les comportements contreviennent aux règles de la plateforme sous l'angle d'une pratique particulière et définie** : fausses déclarations, engagement inauthentique, usurpation d'identité, spams, vandalisme… (**toutes sauf** Bing[65]). Ex. : -> vidéos truquées hyperréalistes (ou *deepfakes*) créées grâce à des technologies d'intelligence artificielle : « *vidéos manipulées et trompeuses* » (Facebook) ; « *médias synthétiques et manipulés* (Twitter) ; -> Comptes « faux-nez » (*sockpuppet*) (Wikipédia) : détention par un utilisateur de plusieurs comptes utilisés à des fins contraires à l'intérêt du service (ex. : s'encourager soi-même dans un débat, voter plusieurs fois) ; -> pratique du *flood*[66] ou du *up*[67] sur un forum (Jeuxvideo.com) ; -> exploitation des *data voids*[68] (vides de données) (Microsoft).
Opérations ou campagnes d'influence coordonnées	Facebook distingue deux types : - les comportements inauthentiques coordonnés dans le cadre de campagnes nationales non gouvernementales[69] : groupes de comptes trompant sur leur identité et activités, s'appuyant sur de faux comptes ; - les ingérences étrangères ou gouvernementales[70] : campagnes d'influence menées au nom d'une entité gouvernementale ou par un acteur étranger. Google (opérations d'influence coordonnées), Twitter (activités inauthentiques coordonnées). Doctissimo, Dailymotion, Snapchat, Yahoo, Jeuxvideo.com n'ont pas eu connaissance d'opérations de ce type sur leur service.

MOYENS DE LUTTE CONTRE LA DIFFUSION MASSIVE DE FAUSSES INFORMATIONS	
Les pratiques de diffusion massive de fausses informations ne font pas toujours l'objet d'un traitement en tant que telles, mais traités dans le cadre de la lutte contre les contenus faux.	
Moyens d'identification des comptes propageant massivement de	-> Moyens exclusivement humains : signalements et examens par les modérateurs (Dailymotion ; Jeuxvideo.com, Doctissimo pour les *floods,* Twitter pour les *deepfakes*) ; -> Moyens automatiques et humains *(*Facebook, Google, LinkedIn, Microsoft, Yahoo, Twitter[71], Jeuxvideo.com[72], Wikipédia).

[65] Bing estime ne pas être concerné dans la mesure où il n'est pas possible de créer un compte sur son service.
[66] Multi-publication de contenus.
[67] Publication de messages intempestifs dans un fil de discussion pour le faire remonter dans la liste des fils.
[68] Termes dont la recherche aboutit à un nombre limité de données pertinentes et d'autorité, ce qui rend facilement accessibles les données problématiques associées.
[69] CIB pour *Coordinated Inauthentic Behavior*.
[70] FGI pour *Foreign or Government Interference*.

fausses informations	Création ou mise à jour d'algorithmes de détection en 2020 : -> Facebook : mise à jour de l'outil de détection proactive pour lutter contre les commentaires de masse et détecter les spams ; -> Twitter : amélioration des systèmes de détection de lutte contre l'automatisation malveillante ; -> Wikipédia : algorithme de détection des comptes « faux-nez » abusifs[73].
Critères de détection et traitement des comptes propageant massivement des fausses informations	**<u>Critères fondés sur le comportement de l'utilisateur</u>** -> création abusive de comptes par le même utilisateur (Google, LinkedIn, Twitter, Wikipédia) ; -> faisceau d'indices : incitation au clic, titre trompeur (Dailymotion) ; utilisation d'informations de profil volées, publication de hashtags identiques, amplification artificielle coordonnée (Twitter)[74] ; présence persistante d'un fil de discussion en page d'accueil, contenu diffusé plusieurs fois par un *bot* (Jeuxvideo.com) ; republication d'une fausse information déjà supprimée à plusieurs reprises (Doctissimo) ; -> usurpation d'identité (Google) ; -> publication massive par des « comptes populaires[75] » (Snapchat) ; -> publication répétée de fausses informations par un compte (LinkedIn). -> analyse automatisée du compte à l'aune de critères successifs (« entonnoir de défense ») et du contenu via des modèles de *machine learning*, des « classificateurs » et du *hash matching*[76] (LinkedIn).
Moyens d'identification des opérations d'influence coordonnées	Enquêtes menées : -> Facebook : observation du comportement des acteurs ; -> Google : groupe d'analyse des menaces (*Threat Analysis Group*, TAG) surveillant des groupes d'acteurs malveillants identifiés ; -> LinkedIn : équipe dédiée sur les comptes soutenus par des entités étatiques ou institutionnelles ; -> Twitter : documentation et archivage des comptes liés à des opérations d'influence, en collaboration avec journalistes, chercheurs et autorités.
Moyens d'identification des *deepfakes*	Critères suivis par Facebook : - contenu édité ou raccourci pour des raisons autres que de clarté ou de transparence et peut laisser croire que le sujet de la vidéo tient des propos qu'il n'a jamais prononcés ; - contenu produit d'une IA qui fusionne, remplace ou superpose des contenus dans une vidéo qui semble vraie ; - exclusion des contenus parodiques ou satiriques.

[71] Détection des comptes automatisés par reCAPTCHA et demande de réinitialisation du mot de passe.
[72] Tout nouvel utilisateur doit remplir un *captcha* lors de la saisie de ses premiers messages.
[73] Une fois le compte identifié comme suspect (selon la qualité de ses contributions, il est analysé à l'aune de plusieurs indicateurs (similarités des contributions, ton promotionnel, correspondance des adresses IP). Les comptes bloqués sont listés sur l'encyclopédie et référencés dans une API.
[74] Twitter combine les moyens humains et automatisés pour les comportements suspects s'agissant des comptes (vol de profil, coordination entre comptes), de l'engagement (inflation des métriques, vente ou achat de Tweets) et de l'utilisation des fonctionnalités (ex. : *hashtags*).
[75] Comptes qui publient un grand nombre de *Stories* publiques.
[76] Recherche de correspondance d'empreintes.

	Critères suivis par Twitter : - contenu synthétique ou manipulé ; - contenu partagé de manière trompeuse (analyse du contexte) ; - contenu susceptible d'impacter la sécurité publique ou de causer un préjudice grave. Utilisation d'outils d'IA et partenariats avec la recherche, la société civile et l'industrie technologique (Microsoft avec Facebook, Google et Wikipédia).

TRAITEMENT DES PRATIQUES ET DES COMPTES PROPAGEANT MASSIVEMENT DE FAUSSES INFORMATIONS	
Mise en œuvre de la décision : humaine ou automatisée	Décision de clôture d'un compte par une équipe humaine (Jeuxvideo.com, Snapchat, LinkedIn, Dailymotion, Doctissimo). Procédés utilisant l'IA et l'automatisation (Google, Wikipédia, Facebook) : - Facebook : détection et suppression d'un faux compte fondées sur les signalements et/ou les systèmes automatiques de détection ; - Wikipédia : suppression des contenus (tout contributeur) ; blocage en écriture de comptes d'utilisateurs (administrateurs et contributeurs) ; annulation de contributions automatiquement classées dans la catégorie vandalisme (programmes informatiques tels que « ClueBotNG »).
Moyens alloués	Pas d'information sur les moyens humains et financiers spécifiquement alloués.
Mesures pouvant être prises en cas de propagation massive de fausses informations	**Mesures sur le contenu** - modification et protection en écriture[77] du contenu (Wikipédia) ; - suppression (Snapchat, Dailymotion, Facebook, Wikipédia, Doctissimo). **Mesures sur le compte** - blocage (Google, Twitter[78], Facebook, Microsoft, Yahoo, Jeuxvideo.com, Wikipédia, Doctissimo) ; - suppression (Google, Microsoft, Snapchat, Twitter, LinkedIn, Dailymotion, Jeuxvideo.com, Facebook, Doctissimo) ; - restriction (*Yahoo*, *LinkedIn*) ; - avertissement (*Jeuxvideo.com*) ; - exclusion temporaire ou définitive du service (*Jeuxvideo.com*) ; - blocage avant ou lors de la création du compte (Facebook) ; - réduction de la visibilité de la page et de ses publications et perte de capacité d'enregistrement en tant que page d'actualités (Facebook).
Nombre de comptes et contenus	Facebook : les faux comptes (hors ceux bloqués en amont) représentent environ 5 % des utilisateurs actifs mensuels au 4e trimestre 2020 ; 5,8 milliards ont été désactivés en 2020.

[77] Protection totale ou une semi-protection de pages signalée par une icône « cadenas » en haut de la page.
[78] Verrouillage d'un compte (fonctions tweets, retweets et *like* interdites) en cas de détection automatique d'une activité inhabituelle.

détectés et traités (niveau mondial, sauf précision)	YouTube : au 4e trimestre 2020, pour violation des règles sur le spam, les pratiques trompeuses et les escroqueries : suppression de 1,4 millions de vidéos (15,5 % des vidéos supprimées) et d'environ 1,8 millions de chaînes. Microsoft Advertising : suspension de près de 300 000 comptes et suppression de 1,6 milliard d'annonces et 270 000 sites.
Mesures économiques (restrictions ou incitations)	Google : procédures contre les tentatives de monétisation d'acteurs malveillants.
Mesures contre les *deepfakes*	Facebook, Twitter : suppression (si menace).
Mesures prises à l'encontre des comptes liés à des opérations d'influence coordonnées	Facebook : - comportement inauthentique coordonné : suppression des comptes, pages et groupes inauthentiques et authentiques directement impliqués ; - ingérence étrangère ou gouvernementale : suppression des données liées à l'opération et aux personnes et organisations qui la soutiennent ; - dans tous les cas : surveillance des réseaux précédemment supprimés ; échange d'informations avec des chercheurs indépendants, des organismes gouvernementaux et partenaires industriels. Données chiffrées (monde, décembre 2020) : suppression de 1957 comptes Facebook, de 707 comptes Instagram, de 156 pages et de 727 groupes. Google : - suppression de chaînes YouTube et blogs des acteurs impliqués, option de monétisation réduite) ; - en cas de tentative de piratage et d'hameçonnage : avertissement des utilisateurs dont les comptes sont la cible de pirates soutenus par le gouvernement (en avril 2020 : 1755 avertissements) ; Programme de Protection Avancée (APP) conçu pour les comptes à haut risque. Twitter : publication des archives complètes des opérations d'information soutenues par des entités étatiques (en 2020, 54 254 comptes ont été archivés).
Mesures de lutte contre l'accélération et la viralité	Facebook : détection et traitement des spams avant ou après leur création. Google : lutte contre les tentatives de manipulation artificielle du taux d'engagement (ex. : via la fonctionnalité « je n'aime pas » sur les vidéos). Twitter : interdiction du gonflage artificiel de l'engagement[79].
Recettes générées pour la plateforme par ces comptes et	**Aucune recette** : Facebook (du fait de la suppression très rapide) ; Google (du fait de ses règles de monétisation) ; LinkedIn (pas de rémunération des influenceurs) ; Jeuxvideo.com.

[79] Ex. : vente ou achat de Tweets ou de comptes gonflés, de *followers* ou d'engagements (Retweets, *likes*, mentions, votes sur des sondages), utilisation ou promotion de services tiers pour gagner des *followers*, etc.

recettes reversées à ces comptes	**Absence de données** - pas de réponse (Twitter, Yahoo). - pas concerné (Wikipédia, Bing, Doctissimo, Snapchat).

INFORMATION DES UTILISATEURS	
Information aux utilisateurs sur les mesures de détection et de traitement de tels comptes	Dans les centres d'aide, charte d'utilisation ou Standards de la communauté (Dailymotion, Yahoo, Facebook, Google, Snapchat, Twitter, Wikipédia, Jeuxvideo.com, Doctissimo) ; Dans les rapports de transparence et les archives (Twitter, LinkedIn) ou rapports mensuels sur les comportements inauthentiques coordonnés (Facebook, Google).
Information aux utilisateurs sur les risques liés à la pratique de création de tels comptes	Accessibilité et niveau de détail : - langage simple et clair : illustrations des comportements interdits (*Twitter*) ; synthèse des pages légales (*Dailymotion*) ; - dans des espaces techniques (Blog d'Ingénierie) et décrites à un niveau de détail moindre pour les utilisateurs pour éviter le contournement (*LinkedIn*). Informations délivrées aux utilisateurs : - interdictions (ex. : *Wikipédia* : compte « faux-nez », vandalisme) ; - risques encourus (*Dailymotion, Facebook, Google, Snapchat, Jeuxvideo.com, Yahoo, LinkedIn, Doctissimo*) ; - modalités de clôture d'un compte (Dailymotion).

MESURES EN CAS DE SITUATIONS EXCEPTIONNELLES (mesures prises pendant les périodes de crise sanitaire et d'élections)	
Impact des élections	Pendant les élections présidentielles américaines de 2020 : - groupe interne de lutte contre la manipulation de l'information : détection de 18 évènements non liés à des opérations d'acteurs coordonnés ou subventionnés par des États (Wikipédia) ; - mesures de lutte contre la désinformation (Facebook, Google[80]).
Impact de la crise sanitaire mondiale	Mesures spécifiques contre des attaques de groupes (Google) : - lutte contre les tentatives de piratage et d'hameçonnage d'acteurs malveillants commerciaux soutenus par des gouvernements[81] ; - suspension de 1800 comptes d'annonceurs établis en UE, dont 1500 en France, pour tentative de contournement, notamment pour des annonces liées à la Covid-19.

[80] *Google* a contrôlé plus de 5400 nouveaux annonceurs électoraux en 2020.
[81] Exemple : **sociétés de « *hack-for-hire* »** créant des comptes Gmail se faisant passer pour l'OMS et ciblant des dirigeants d'entreprises de services financiers, de conseil et de santé dans plusieurs pays.

5. Mesures de lutte contre les fausses informations en matière de communications commerciales et de promotion des contenus d'information se rattachant à un débat d'intérêt général

RESTRICTIONS DE LA PUBLICATION DE COMMUNICATIONS COMMERCIALES SUR LA PLATEFORME	
Refus des contenus promus d'information se rattachant à un débat d'intérêt général	**De manière permanente** (Dailymotion ; Jeuxvideo.com ; LinkedIn ; Microsoft ; Doctissimo) ; interdiction des publicités politiques uniquement (Twitter). **De manière temporaire**, principalement pendant les périodes électorales (Google[82] ; Snapchat ; Yahoo ; Facebook).

SIGNALEMENT DES ANNONCES PUBLICITAIRES ET CONTENUS SPONSORISÉS [83]	
Processus de signalement spécifique	Processus de **signalement spécifique pour « fausse information »** (*Facebook ; Microsoft ; Snapchat*),
Suivi de l'état du signalement par l'utilisateur	**Peu d'informations sur ce qu'il advient des contenus signalés et non encore traités** sauf **:** - LinkedIn : une publicité est mise en quarantaine après plus d'un signalement afin que les équipes de modération l'étudient ; - Facebook, LinkedIn, Twitter : suivi du signalement possible par l'auteur du signalement.
Traitement (automatisé ou humain)	- LinkedIn : les publicités ayant reçu plus d'un signalement sont suspendues puis **examinées manuellement** ; - Google, Facebook, Microsoft et Twitter : vérification **par des algorithmes et des moyens humains** (répartition des tâches non décrite) ; - Dailymotion : intervention manuelle de l'équipe dédiée à la sécurité des marques.

VISIBILITÉ DES ANNONCES PUBLICITAIRES ET CONTENUS SPONSORISÉS	
Modalité de vente des espaces publicitaires	Les systèmes décrits consistent généralement en un modèle fermé reposant sur des outils propres à la plateforme : - système d'enchères en temps réel géré automatiquement (Twitter ; Facebook ; LinkedIn ; Microsoft Advertising ; Doctissimo) ; - plateforme de publicité en libre-service (Snapchat) ; - système de gré à gré, où annonceurs et agences achètent directement l'espace à une régie, associé à un système de publicité programmatique pour vendre les espaces restants (Dailymotion) ; - facturation au coût par clic ou CPM (Google).
Moyens permettant de gagner en visibilité	Ciblage de l'audience par le biais de différents critères (Dailymotion ; Facebook ; LinkedIn ; Microsoft Advertising ; Snapchat ; Twitter) : format, intérêt, âge, localisation, mot-clef, sexe, type d'appareil utilisé, etc. (variables selon les plateformes).

[82] Sauf pour les « *annonces d'information sur les élections diffusées par les organes officiels de communication du Gouvernement* ».
[83] A lire en parallèle de la partie sur les signalements des contenus en général (p.X).

	Pertinence de la visibilité de l'annonce publicitaire parfois obtenue par le biais d'algorithmes (ex : Facebook).
Informations chiffrées sur la visibilité et l'importance des flux financiers entourant les annonces publicitaires et contenus sponsorisés	**Très peu d'informations chiffrées – ou déclarées comme couvertes par le secret des affaires –** relatives au volume d'annonces publicitaires et contenus sponsorisés en lien avec une infox, aux revenus générées par celles-ci et à leur visibilité. LinkedIn précise procéder à une vérification des publicités avant publication, empêchant ainsi ce type de contenus d'apparaître. Twitter souligne la difficulté d'isoler ces données.

DÉTECTION DES ANNONCES PUBLICITAIRES ET DES CONTENUS SPONSORISÉS PORTEURS D'INFOX	
Vérification des annonceurs[84] et des publicités	Création d'un compte et/ou procédure spécifique (ex : Facebook ; Google ; Dailymotion ; LinkedIn ; Snapchat ; Twitter). Informations complémentaires demandées pour vérifier l'existence d'une personne morale ou de son identité (Facebook ; Google ; Twitter).
Mesures de détection	Mesures de contrôles préalables des annonces, notamment par des moyens humains (ex : Dailymotion[85]). Utilisation d'algorithmes de détection (ex : Google ; Twitter) parfois couplée à un examen humain (ex : Snapchat ; LinkedIn). Détection par des partenaires de vérification des faits (ex : Facebook[86]).
Moyens de coopération mis en place avec d'autres opérateurs ou des organismes tiers	Facebook coopère avec des institutions publiques, des autorités judiciaires ou de police, des acteurs privés et des universitaires et chercheurs. Snapchat travaille avec des tiers vérificateurs de confiance (trusted flaggers) pour le signalement des publicités porteuses de fausses informations. Doctissimo fait appel à une technologie externe de lutte contre la fraude publicitaire (« Confiant »).

[84] En général, et non uniquement les annonceurs politiques évoqués dans le tableau « promotion des contenus d'information se rattachant à un débat d'intérêt général ».
[85] Dans les cas de publicité en gré à gré.
[86] Voir partie sur les vérificateurs de faits (p.X)

MESURES LIÉES À LA SÉCURITÉ DES MARQUES	
Mesures de sécurité des marques	- Examen des signaux contextuels des produits publicitaires ; - interdiction des publicités superposées ou en pop-up (Google ; Snapchat ; Twitter ; Dailymotion) ; - possibilité de sélectionner les placements sur lesquels l'annonceur souhaite que ses annonces apparaissent (Facebook) ; - possibilité pour l'annonceur de voir où sa publicité est apparue (Facebook) ; - interdiction des publicités incluant un contenu démenti par des partenaires vérificateurs de faits (ex : Facebook).
Moyens dédiés à l'exécution de ces mesures	- équipe dédiée au sein de l'entreprise (ex : Snapchat) ; - équipe interne et partenariats avec des tiers (ex : Dailymotion ; Facebook ; Twitter)
Évaluation de l'impact de ces mesures	Très peu d'informations sur ce point.
Information des annonceurs	Notification des annonceurs et possibilité de recours si la plateforme rejette l'annonce en considérant qu'elle véhicule une infox (Facebook). Compensation ou possibilité de recours pour les annonceurs dont l'annonce a été attachée à un contenu véhiculant une infox (ex : Dailymotion)
Coopération	Coopérations avec les marques, des représentants du monde publicitaire ou des organisations expertes (Dailymotion ; Facebook ; Twitter)[87].

MESURES LIÉES AUX CONTENUS DES UTILISATEURS RÉALISÉS EN PARTENARIAT AVEC DES TIERS	
Moyens de lutte contre ceux de ces contenus porteurs d'infox	Pas de mesure particulière pour ces contenus, non gérés par les répondants. La majorité leur applique les mêmes règles qu'aux annonces publicitaires et contenus sponsorisés.

PROMOTION DES CONTENUS D'INFORMATION SE RATTACHANT À UN DÉBAT D'INTÉRÊT GÉNÉRAL	
Définition	**Pas de définition des « contenus d'information se rattachant à un débat d'intérêt général » stricto sensu, souvent rapprochés d'autres notions** (définies dans les politiques publicitaires de la plateforme) : - « publicités à caractère politique et militant » (Snapchat) ; - « publicités défendant une question législative d'importance nationale, publicités faisant référence à une élection, un parti politique ou un candidat clairement identifié » (Twitter).
Identification des annonceurs	Afin notamment d'éviter l'influence de puissances étrangères, **processus d'identification particuliers pour les annonceurs politiques** : autorisation dans

[87] Depuis 2020, Dailymotion et Facebook sont titulaires de la certification *Trustworthy Accountability Group* (TAG), Twitter et Facebook sont membres de la Global Alliance for Responsible Media, Twitter coopère avec DoubleVerify, Integral Ad Science et le Media Ratings Council et Dailymotion travaille avec l'association Adtech IAB.

politiques	un seul pays, obligation de fournir un document d'identité pour le pays ciblé… (Facebook ; Google[88] ; YouTube; Snapchat[89] ; Twitter[90]).
Restrictions et examen particulier	**Restrictions de ciblage** (Twitter[91] ; Facebook[92]).

MISE EN PLACE DE BIBLIOTHEQUES PUBLICITAIRES	
Mise en place de bibliothèques publicitaires publiques	**Bases de données publicitaires publiques** (Facebook ; Google ; Twitter ; Snapchat), qui peuvent prendre **différentes formes** : bases de données avec moteurs de recherche, API, centres de transparence, rapports, pages… LinkedIn : l'**onglet « publicités » sur les pages Entreprise** (vers lesquelles le contenu sponsorisé renvoie) permet de voir les contenus sponsorisés actuels ou passés pour lesquels LinkedIn a été payé par le propriétaire de la page ou une personne affiliée. Pas d'information fournie (Yahoo ; Microsoft[93]) ; pas de bibliothèque publicitaire (Doctissimo).
Contenus	- Facebook : bibliothèque publicitaire publique et consultable sans compte, incluant toutes les publicités « actives et en cours » (sauf les publicités politiques qui sont conservées 7 ans) ; - Snapchat : bibliothèque des publicités politiques et militantes.

MESURES PRISES DANS LE CADRE DE LA CRISE DE LA COVID-19	
Initiatives prises dans le cadre de la crise de la Covid-19	Mise en place de coopérations avec, notamment, des organismes de santé ou des organismes gouvernementaux pour promouvoir les informations faisant autorité (ex : Facebook ; Google). Mise en place de mesures empêchant la monétisation des contenus pouvant promouvoir des pratiques dangereuses pour la santé (ex : Google). Extensions de politiques existantes afin de couvrir les contenus en lien avec la Covid-19 et/ou mise en place de restrictions temporaires pour ces contenus (ex :; LinkedIn ; Twitter).

[88] Depuis début 2021 en France : processus de validation des annonceurs étendu à toutes les annonces (plus uniquement aux annonceurs politiques) ; fonctionnalité permettant d'avoir plus d'informations sur l'annonceur des publicités diffusées sur Google Ads.
[89] Une publicité politique ne peut pas être payée par des personnes ou entités qui ne résident pas dans le pays où elle sera diffusée.
[90] Twitter limite la publicité et exige la certification des annonceurs pour les publicités « *qui éduquent, sensibilisent ou appellent les gens à agir en rapport avec l'engagement civique, la croissance économique, la gestion de l'environnement ou les causes d'équité sociale* ».
[91] Pour les publicités dédiées à une cause, les possibilités de ciblage se limitent au ciblage géographique, par mots-clés et par intérêts non politiques. Aucun autre type de ciblage n'est autorisé, y compris les audiences personnalisées.
[92] A noter que Facebook peut appliquer des restrictions de ciblage à tous les types de contenus publicitaires.
[93] Microsoft précise ne pas avoir de base de données concernant la promotion de contenus d'information se rattachant à un débat d'intérêt général mais ne précise pas ce qu'il en est pour les communications commerciales en général.

6. Éducation aux médias et à l'information et relations avec le monde de la recherche

MESURES PERMETTANT D'IDENTIFIER DES CONTENUS PERTINENTS ET FIABLES	
Mesures permettant d'identifier des contenus pertinents et fiables [94]	Facebook : « Index des pages d'actualité » répertoriant toutes les pages publiant des actualités. Dans le cadre de l'épidémie de Covid-19, la plupart des plateformes ont mis en avant des informations officielles ou *fact-checkées* (Facebook, Microsoft, Twitter, LinkedIn, Snapchat, Yahoo, Doctissimo, Google, Bing, Yahoo, Jeuxvideo.com)
MESURES PERMETTANT DE FORMER À L'USAGE DES PLATEFORMES ET À L'IDENTIFICATION DES CONTENUS	
Initiatives permettant de former les publics à l'utilisation des plateformes	Formations menées en partenariat. Ex. : • Google : collaboration avec l'Observatoire pour la parentalité et l'éducation au numérique, Génération Numérique ou encore l'association « Les Petits Débrouillards »[95] ; • Facebook : portail pour les jeunes ; partenariat avec Freeformers et plus de 20 ONGs nationales[96]. Campagne d'EMI dédiées dans le cadre de l'épidémie de Covid-19. Ex. : • Twitter : campagne lancée en partenariat avec l'Unesco et la Commission européenne, « *#ThinkBeforeSharing* » ; • Google : « *Family link* » ; • Facebook : fonds pour le civisme en ligne qui consacre un million d'euros, en France, aux actions en matière d'EMI.
Initiatives permettant aux publics de développer leur esprit critique	Initiatives d'EMI sur le service. Ex. : • Facebook : campagne « Trois questions pour aider à éradiquer les *fake news* » consacrée à la lutte contre les fausses informations liées au Covid-19 ; • Facebook *:* l'élargissement de l'initiative de *redirect* dans le cadre de la lutte contre les organisations ou mouvements liées à la violence. NB : aux États-Unis, à l'occasion des élections, lancement par Microsoft de campagnes visant à renforcer la culture du numérique des citoyens et à les sensibiliser à l'impact des médias sur les démocraties • « *Spot the Deepfake* », • soutien à une campagne lancée en collaboration avec la *Radio Television Digital News Association*, le *Trust Project* et le *Center for an Informed Public and Accelerating Social Transformation Program* de l'Université de Washington qui incite les citoyens à prendre le temps de réfléchir à ce qu'il leur est donné à voir sur Internet.

[94] Se référer à la partie 3 du bilan sur la promotion des contenus issus d'entreprises et d'agences de presse et de services de communication audiovisuelle (p.X).
[95] Formation suivie par 61 000 jeunes âgés de 6 à 14 ans en 2020.
[96] Programme de formation dans sept pays européens, dont la France, pour développer la confiance des utilisateurs qui a été suivi par 75 000 personnes.

Initiatives à destination des professionnels de l'information	Mise en place de mesures pour soutenir les rédactions pendant la crise de Covid-19. Ex. : Google, avec le soutien de Google News Initiative, les plateformes Big Local News et Pitch Interactive de l'Université de Stanford : outil de cartographie des cas de Covid-19 à l'échelle mondiale qui permet aux journalistes d'incruster des représentations actualisées de l'évolution de la pandémie dans leurs articles en ligne, etc. Développement de formations consacrées à la vérification d'information, notamment en période d'élections. Ex : Google et sa *Google News Lab*.
AIDES ET PARTENARIATS AVEC LE MONDE DE LA RECHERCHE	
Financements	Financement de chaires universitaires, de colloques et groupes de recherche. Ex. : partenariat de Facebook avec l'École supérieure de journalisme de Lille pour mettre en place des outils et des programmes efficaces d'éducation au numérique. Financement d'initiatives pour améliorer le modèle économique du journalisme en ligne. Ex. : projet « Accelerated Mobile Pages » Google.
Partage de données	Accès facilité aux algorithmes et aux API. Ex. : Twitter : ouverture d'un point d'accès spécifique au Covid-19 sur son API et lancement d'une mise à jour pour répondre aux besoins des chercheurs. Réflexions engagées pour parvenir à déterminer des protocoles permettant de mieux partager les données. Ex. : • Facebook : plateforme Facebook Open Research & Transparency; • Twitter Academic Resarch Product Track[97].
Initiatives de recherche spécifiques sur les fausses informations	Collaborations avec des organismes de recherche. Ex. : • Google, auprès du Fonds Européen pour les médias et l'information ; • Microsoft, avec Princeton sur un projet intitulé « *Trends in online foreign influence operations* » et avec l'Oxford Internet Institute pour le développement d'outils ; • Wikipédia, avec la Federal University of Espírito Santo et la Teléfonica Research sur la manipulation de l'information ; avec la New York University sur les comptes « faux-nez ». Lutte contre les *deepfakes*. Ex. : • le partenariat sur l'AI (Facebook, Microsoft, Wikipédia, Google) ; • Facebook : partenariats avec le MIT, Berkeley ou encore Cornell Tech pour son *Deepfake Detection Challenge* ; • Google : mise à disposition de données dans le cadre de l'ASVsproofchallenge ; la contribution aux jeux de données en open source.

[97] Mise à jour permettant aux chercheurs d'obtenir un accès gratuit à l'historique complet des conversations publiques, des niveaux d'accès plus élevés et gratuits à la plateforme de développement Twitter et des capacités de filtrage plus précises.