Arcom
Le régulateur de la communication audiovisuelle et numérique

Juillet 2023

L'essentiel

Bilan et perspectives de la lutte contre la diffusion de contenus haineux en ligne : *une démarche de responsabilisation des plateformes à consolider dans la perspective du règlement sur les services numériques.*

La loi du 24 août 2021 confortant le respect des principes de la République a renforcé le cadre juridique de la responsabilité des principaux opérateurs de plateforme en ligne actifs sur le territoire français pour lutter contre la diffusion de contenus haineux sur leur service.

Ce cadre national visait à anticiper le **règlement européen sur les services numériques** (RSN), adopté le 19 octobre 2022, dont il s'inspire fortement. Il disparaîtra début 2024 pour lui laisser la place, en vertu du principe d'harmonisation maximale. Pour les plus grandes plateformes en ligne, c'est dès le 25 août 2023 que le RSN entrera en application, sous le contrôle de la Commission européenne.

Le bilan dressé par l'Arcom permet d'expérimenter ce cadre. **Sur la base d'observations et des déclarations des opérateurs en réponse à un questionnaire *ad hoc***, il fait état de l'amélioration des outils et des procédures mis en place pour lutter contre les usages abusifs des principales plateformes en ligne.

Un effort de transparence à consolider

Les opérateurs interrogés ont, dans l'ensemble[1], fait preuve d'une transparence renforcée. Les déclarations apportent des informations chiffrées, inédites et de qualité, notamment sur le nombre et l'origine des signalements de contenus à caractère haineux, les moyens humains de modération et leur collaboration avec les autorités publiques et les signaleurs de confiance.

En particulier, certains acteurs, tels que *Dailymotion* ou *Meta*, ont développé des outils et indicateurs pertinents pour **objectiver la prévalence** des contenus illicites sur leurs services. Ces efforts contrastent avec les insuffisances relevées s'agissant de services tels que *Twitter* ou *Bing* (*Microsoft*).

Cette démarche de transparence reflète une **volonté tangible de préparer la mise en conformité avec le RSN,** dans le cadre duquel les VLOPSEs[2] devront faire preuve d'une transparence renforcée et auront l'obligation d'évaluer les risques systémiques liés à leur plateforme et de prendre des mesures pour les atténuer.

Les difficultés rencontrées par certaines plateforme[3] pour rassembler et transmettre des données détaillées dans le temps imparti témoignent de la nécessité de déployer des moyens d'analyse et de *reporting* à cette fin. L'effort de transparence doit également être poursuivi s'agissant du nombre et de la langue de travail des modérateurs chargés de lutter contre la dissémination des contenus haineux en ligne, cette information d'intérêt général étant encore trop souvent l'objet de demandes de confidentialité, au détriment de l'information du public.

1 *Pinterest* fait figure d'exception.
2 Très grandes plateformes en ligne et très grands moteurs de recherche, dépassant 45 millions de destinataires actifs mensuels dans l'Union européenne, et désignés comme tels par la Commission européenne.
3 Ainsi *Google* ou *TikTok*.

- Tableau sur les moyens humains de modération des services

*Tableau portant sur les **moyens humains de modération** des services*[4]	Nombre de personnes affectées au traitement des signalements et des recours provenant des utilisateurs de la version française du service en matière de haine en ligne	Localisation des personnes affectées au traitement des signalements et des recours provenant des utilisateurs de la version française du service en matière de haine en ligne	Langue(s) de travail des personnes affectées au traitement des signalements et des recours provenant des utilisateurs de la version française du service en matière de haine en ligne
Bing	*« Microsoft dispose d'une variété de cadres de modération de contenu avec des centaines d'employés à temps plein, des employés de fournisseurs et d'autres employés occasionnels »*	*« Situés dans de nombreux endroits à travers le monde »*	*« Bing s'assure que les équipes de modération de contenu chargées d'examiner les réclamations en français disposent d'un personnel adéquat pour prendre en charge la langue française. »*
Dailymotion	Moyenne d'une trentaine de personnes	Union européenne (France et Roumanie)	*« Grande variété linguistique »*
Google Search	Donnée non communiquée	Donnée non communiquée	Donnée non communiquée
LinkedIn	*« Centaines de réviseurs de contenu […] dont environ 29 spécialistes de la langue française »* (ex. : fin 2022, 278 employés à temps plein basés dans la région Europe, Moyen-Orient et Afrique et des contractants)	*« Situés dans le monde entier »*	*« Avec une expertise linguistique locale »*
Meta (Instagram et Facebook)	**confidentiel**	**confidentiel**	*« Incluent le français, l'anglais, l'allemand et le turc. »*
Pinterest	**confidentiel**	**confidentiel**	**confidentiel**
Snapchat	**confidentiel**	**confidentiel**	**confidentiel**
TikTok	**confidentiel**	**confidentiel**	**confidentiel**
Twitter	*« 149 personnes, dont des francophones »*	*Union européenne (« et basées en Irlande et au Portugal »)*	*« dont des francophones »*
Wikipédia	*« Une équipe géographiquement répartie d'environ 45 à 50 personnes était disponible […] la plupart des dossiers de cette nature sont traités par 2 à 6 personnes (environ) ».*	*« Équipe géographiquement répartie »*	*« Intrinsèquement multilingue »*
Portail Yahoo	Donnée non communiquée	Donnée non communiquée	*« en ce compris des juristes francophones si nécessaire »*
YouTube	Donnée non communiquée	Donnée non communiquée	Donnée non communiquée

Un travail à poursuivre en matière de clarté et d'accessibilité des conditions générales (CG)

Le degré d'accessibilité des CG, depuis la page principale des services, apparaît satisfaisant via un navigateur web sur ordinateur (3 clics maximum). Y accéder est cependant souvent légèrement plus complexe sur mobile, via un navigateur web, avec un nombre maximal de 4 clics requis. Leur accessibilité est encore amoindrie sur l'application de certains services alors que c'est, pour la grande majorité d'entre eux, le premier mode d'accès (5 clics maximum).

Au regard de la nature même de ce type de contenus, proposer des CG intelligibles est à la fois un défi et une nécessité. Les CG des services observés sont toutes disponibles intégralement en français, à l'exception de celles de *LinkedIn* qui proposent certaines sections uniquement en anglais. Néanmoins, pour plusieurs d'entre eux, la mise en page des CG pourrait être optimisée afin d'en favoriser l'intelligibilité. Une bonne pratique, anticipant le RSN, mise en place par certains opérateurs consiste à fournir des synthèses de chaque section et à les mettre en avant visuellement au sein des CG.

[4] Informations valant au mois de décembre 2022.

De même, sur la majorité des services observés, la possibilité d'accéder à un système interne de traitement des réclamations d'une décision de modération est mentionnée dans les CG ou dans les règles communautaires. Cependant, trois plateformes[5] se singularisent par l'absence de mention explicite de cette faculté au sein de leurs CG.

Le signalement des contenus et la contestation des décisions de modération

L'Autorité constate que **les opérateurs ont globalement mis en œuvre des moyens pour favoriser le signalement des contenus illicites sur leurs plateformes** et pour impliquer leurs utilisateurs. À l'exception de quelques-uns[6], ils permettent également aux utilisateurs non connectés à un compte de signaler un contenu illicite. Certains services[7] proposent un formulaire *ad hoc* permettant de signaler un contenu à caractère haineux au titre l'article 6-4 de la loi du 21 juin 2004 pour la confiance dans l'économie numérique (LCEN)[8] .

Cependant, sur de nombreux services, **l'accès au dispositif de signalement est conditionné au clic sur un bouton dont l'intitulé est peu explicite** (ex. : « … »).

Ainsi, l'Arcom invite les plateformes à demeurer particulièrement attentives à l'accessibilité de ces dispositifs ainsi qu'à leur intelligibilité.

- Tableau portant sur les signalements émis par les utilisateurs des services[9]

*Tableau portant sur les signalements émis par les **utilisateurs** des services*	Nombre déclaré de signalements de contenus haineux au sens de l'article 6-4 de la LCEN reçus en 2022	Nombre d'actions de modération portant sur des contenus haineux en 2022	Part des actions de modération prises après décision humaine
Bing	0	0	Donnée non pertinente
Dailymotion	387	490[10]	100 %
Google Search	Donnée non pertinente	Donnée non communiquée	100 %
LinkedIn	**confidentiel**	3 064	**confidentiel**
Meta (*Instagram* et *Facebook*)	12 424	4 807[11]	100 %
Pinterest	**confidentiel**	**confidentiel**	**confidentiel**
Snapchat	24 184	5 637[12]	100 %
TikTok	395 302	152 628	37,6 % des suppressions fondées sur une violation des Règles Communautaires, 100 % des géoblocages et suppressions après signalements effectués via le dispositif « Signaler un contenu haineux en France »
Twitter	1 159 206 sur 3 405 570 signalements[13] reçus au total	Mesures prises 289 169 éléments (taux d'action : 24,95 %)	34 % du total des signalements[14]
Wikipédia	0	0	*« Non applicable »*
Portail Yahoo	*« 0 (N/A)»*	0	Donnée non pertinente
YouTube	Donnée non pertinente	Donnée non communiquée	100 %

[5] *Pinterest*, *Snapchat* et *Bing*.
[6] *Twitter, Pinterest* (sauf via un formulaire ad hoc) et *LinkedIn* ne permettent pas à leurs utilisateurs non connectés de signaler un contenu illicite.
[7] Ceux de *Google Search*, *YouTube*, *TikTok, Facebook* et *Instagram* sont accessibles à toute personne ; celui de *Pinterest* uniquement aux utilisateurs connectés.
[8] Issu de la loi du 24 août 2021.
[9] L'Arcom invite à analyser ces chiffres avec prudence : ils ne suffisent pas en eux-mêmes à apprécier les efforts des plateformes et doivent être considérés en lien avec la taille de chacune, sa fréquentation, son modèle, l'exhaustivité et la clarté de son dispositif de signalement, etc.
[10] La contribution de l'opérateur précise que plusieurs décisions de modération peuvent s'appliquer à un même contenu.
[11] Parmi ceux-ci, la contribution de l'opérateur précise que 19 contenus ne violaient pas ses Standards de la Communauté mais ont été bloqués en France en raison d'une violation d'une disposition de l'article 6-4 de la LCEN.
[12] L'opérateur indique également avoir pris des mesures à l'encontre de 4 119 comptes uniques pour violation de ses politiques en matière d'incitation à la haine en France.
[13] La contribution de l'opérateur précise que cette notion couvre les signalements d'utilisateurs et les signaux d'apprentissage machine (« *machine learning* »).
[14] Pour rappel : 3 405 570.

La faculté de contestation des décisions de modération, presque partout accessible aux utilisateurs, demeure cependant inégalement saisie par ceux-ci selon les plateformes. De plus, s'il est difficile de tirer un enseignement général de cette situation, le taux d'infirmation des décisions initiales de modération particulièrement élevé pour *TikTok* (40,8 % des appels à la suite d'un signalement via le dispositif dédié aux violations des Règles Communautaires) et *Dailymotion* (44 %) interroge sur la pertinence de l'action de modération initiale sur ces services.

- Tableau portant sur les recours émis par les utilisateurs des services

*Tableau portant sur les recours émis par les **utilisateurs** des services*	Recours internes provenant d'utilisateurs accédant au service depuis la France	Résultats des recours internes
Bing	Donnée non pertinente	Donnée non pertinente
Dailymotion	6,46 % (sur la période juillet-décembre)	56 % de confirmation de la décision initiale (sur la période juillet-décembre)
Google Search	Donnée non communiquée	Donnée non communiquée
LinkedIn	5	100 % de confirmation de la décision initiale
Meta (*Instagram* et *Facebook*)	625	100 % de confirmation de la décision initiale
Pinterest	**confidentiel**	**confidentiel**
Snapchat	65	100 % de confirmation de la décision initiale
TikTok	34 404 (27 770 en application du dispositif dédié aux violations des Règles Communautaires et 6 634 en application du dispositif « Signaler un contenu haineux en France »)	40,8 % d'infirmation de la décision initiale via le dispositif dédié aux violations des Règles Communautaires et 24,66 % via le second
Twitter	11 538 au titre de la LCEN	600 ont fait l'objet d'une action suite à ces appels (taux d'action : 5,20 %)
Wikipédia	*« Non applicable »*	*« Non applicable »*
Portail Yahoo	Pas de recours interne	Pas de retour interne
YouTube	Donnée non communiquée	Donnée non communiquée

La collaboration avec les experts de la lutte contre la haine en ligne : un atout pour les opérateurs

La majorité des opérateurs de plateforme ont établi, parfois de longue date, **des liens de travail avec des signaleurs de confiance**, en France comme dans le reste du monde. Ces relations sont hétérogènes tant en nombre de partenariats noués qu'au regard de la vitalité de ces derniers. L'efficacité de ces partenariats est fonction des moyens dont les signaleurs de confiance disposent et des investissements dont ils bénéficient.

Le RSN reconnaît **le rôle essentiel de ces partenariats** qui contribuent à l'apaisement de l'espace numérique. Ils seront renforcés et dotés d'un cadre juridique clair imposant aux opérateurs de plateforme de **traiter de manière prioritaire** les signalements reçus de leurs partenaires. Opérateurs et signaleurs de confiance seront également liés par des **obligations de transparence réciproque,** auxquelles veilleront les coordinateurs nationaux des services numériques.

- Tableau portant sur la collaboration avec des signaleurs de confiance en France

Tableau portant sur la collaboration des services/opérateurs avec des signaleurs de confiance en France	Nombre déclaré de signaleurs de confiance	Nombre de signalements de contenus haineux déclarés au sens de l'article 6-4 de la LCEN reçus en 2022
Bing	Aucun	Donnée non pertinente
Dailymotion	*« S'est rapproché de différentes associations coutumières des signalements au cours de l'année 2022 pour envisager des modalités partenariales. »*	0
Google Search	« Aucun. »	Donnée non pertinente
LinkedIn	*« N'a pas actuellement de programme formel pour les signaleurs de confiance. »*	Donnée non pertinente
Meta (*Instagram* et *Facebook*)	17 (dont 4 nommés)	0
Pinterest	**confidentiel**	**confidentie**l
Snapchat	4	0[15]
TikTok	11	17
Twitter	5	242 (taux d'action : 66,12 %)
Wikipédia	Aucun	Donnée non pertinente
Portail Yahoo	Aucun	*« 0 (N/A) »*
YouTube	5	Donnée non communiquée

La bonne coopération des opérateurs avec les autorités judiciaires et administratives est un gage de **l'efficacité de la réponse pénale** en matière de diffusion de contenus illicites.

La qualité et la maturité du dialogue entre les opérateurs et les services enquêteurs spécialisés transparaît dans le **taux particulièrement élevé de réponses positives** (**près de 85 %**) aux demandes d'information des autorités françaises aux fins d'identifier l'auteur potentiel d'un contenu haineux en ligne.

L'impératif d'efficacité de ce travail collectif sera encore renforcée par le RSN : les opérateurs seront tenus de répondre avec diligence aux autorités judiciaires ou administratives les enjoignant à agir contre un contenu illicite ou à donner des informations sur son auteur, et **de justifier avec précision les raisons les conduisant le cas échéant à ne pas accéder à la demande.** Ces injonctions, leurs motifs et les réponses qui y seront apportées par les opérateurs feront l'objet de **rapports annuels publics,** contribuant à la transparence de l'action publique en matière de lutte contre les contenus illicites en ligne.

[15] Révision réalisée le 27 juillet 2023 à la suite d'un rectificatif apporté par *Snapchat* à sa déclaration.

- Tableau portant sur les demandes et signalements émis par les autorités publiques françaises

Tableau portant sur les demandes et signalements émis par les **autorités publiques** *françaises*	Nombre de demandes et de signalements de contenus haineux au sens de l'article 6-4 de la LCEN reçus en 2022	Taux d'action de modération portant sur des contenus haineux en 2022	Nombre de transmissions de suspicions d'infraction aux autorités publiques compétentes en 2022
Bing	155 demandes de suppression de contenu	Donnée non communiquée	0
Dailymotion	0	Donnée non pertinente	0
Google (*Search* et *YouTube*)	6 017 demandes de communication de données reçues entre le 1er janvier et le 30 juin	Communication des informations pour 85 % des demandes	0
LinkedIn	0	Donnée non pertinente	0
Meta (*Instagram* et *Facebook*)	0 signalement mais 25 451 demandes de données (dont 12 520 dans le cadre de procédures juridiques et 12 938 demandes de divulgation d'urgence)	Données transmises dans environ 85 % des cas	Donnée non communiquée
Pinterest	**confidentiel**	*« Nous examinons ces demandes avec diligence et transmettons les données pour celles qui répondent aux exigences de la loi et de nos politiques. »*	**confidentiel**
Snapchat	0	Donnée non pertinente	Donnée non communiquée
TikTok	11	Donnée non communiquée	Donnée non communiquée
Twitter	869 demandes de suppression de contenu et 5 032 demandes d'information	43,5 % pour les demandes de suppression et 42,90 % pour les demandes d'information	0
Wikipédia	0	Donnée non pertinente	Donnée non communiquée
Portail Yahoo	0	Donnée non pertinente	0

Méthodologie

Les plateformes visées par la loi du 24 août 2021 sont celles dont la fréquentation dépasse dix millions de visiteurs uniques par mois en France, en moyenne, sur la dernière année civile.

Afin de préparer ce point d'étape, un questionnaire a été envoyé, le 28 avril 2023, aux opérateurs des treize services suivants: *Google* (pour *Google Search* et *YouTube)*, *LinkedIn*, *Meta* (pour *Facebook* et *Instagram*), *Microsoft* (pour *Bing*), *Pinterest*, *Snap, TikTok*, *Twitter*, la *Fondation Wikimédia*, *Yahoo* (pour *Yahoo Search*) et *Dailymotion*. Mis à part *Yahoo Search*, tous ces services sont soit établis en France, soit des VLOPSEs au sens du RSN.

En complément, l'Arcom a procédé à des observations des conditions générales (CG) et dispositifs de signalement sur chaque service pour nourrir son analyse, entre avril et juin 2023, via le navigateur web Google Chrome pour le site internet des services et via les systèmes d'exploitation iOS et Android pour leur application.

Directeur de la publication :
Roch-Olivier Maistre