Arcom

Le régulateur de la communication audiovisuelle et numérique

Lutte contre la diffusion de contenus haineux en ligne

Bilan des moyens mis en œuvre par les plateformes en ligne en 2022 et perspectives

Juillet 2023

Sommaire

Introduction

Le développement des services de la société de l'information a profondément transformé la façon dont des millions d'utilisateurs communiquent et échangent l'information.

Toutefois, l'utilisation accrue de ces services, notamment des plateformes permettant le partage d'informations entre utilisateurs (les réseaux sociaux en premier lieu), a également engendré de nouveaux risques pour la cohésion et le fonctionnement démocratique de notre société, tels que la large diffusion de contenus manifestement illicites, la révélation de phénomènes de mésinformation, et, dans certains cas, de désinformation, ou la mise en évidence d'effets induits préjudiciables tels l'aggravation de problèmes de santé publique, l'accroissement de la conflictualité du débat public en ligne ou la perte de confiance dans nos espaces informationnels. Cette réalité suscite une attente de renforcement de la responsabilité des fournisseurs de services numériques.

Face à l'urgence de la situation générée par la diffusion de contenus illicites en ligne, certains États européens, notamment la France, l'Allemagne et l'Autriche, ont anticipé ce texte en adoptant en amont des premiers cadres législatifs nationaux contraignants.

En France, les dispositions de l'article 42 de la loi confortant le respect des principes de la République du 24 août 2021, qui ont introduit un article 6-4 dans la loi n° 2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique (LCEN), sont venues renforcer la lutte contre les contenus haineux en imposant des obligations procédurales et de moyens, tant humains que technologiques, aux principales plateformes en ligne reçues sur le territoire national. Elle a confié à l'Autorité de régulation de la communication audiovisuelle et numérique (Arcom) la mission de superviser la mise en œuvre de ces obligations. Ce dispositif s'inspire largement de certaines dispositions de la proposition initiale de Règlement sur les services numériques (RSN, ou DSA en anglais) présentée par la Commission européenne en 2020[1]. Ces dispositions nationales doivent s'éteindre au 31 décembre 2023[2].

C'est en réponse à cette attente et en tirant partie de l'expérience acquise des dispositifs nationaux, notamment en France, en Allemagne et en Autriche, que le législateur de l'Union européenne (UE) a adopté le RSN le 19 octobre 2022[3] afin d'établir des règles harmonisées pour un environnement en ligne sûr, prévisible et fiable dans lequel les droits fondamentaux des citoyens européens seront efficacement protégés.

[1] Proposition de règlement du Parlement Européen et du Conseil du 15 décembre 2020 relatif à un marché intérieur des services numériques (Législation sur les services numériques) et modifiant la directive 2000/31/CE.
[2] Il est envisagé, dans le projet de loi visant à sécuriser et réguler l'espace numérique déposé le 10 mai 2023 au Sénat, de prolonger cette échéance jusqu'au 17 février 2024.
[3] Règlement (UE) 2022/2065 du Parlement Européen et du Conseil du 19 octobre 2022 relatif à un marché unique des services numériques

À cette fin, ce règlement réaffirme et renforce le régime de responsabilité limitée à raison des contenus hébergés, *mais* simultanément introduit une série d'obligations nouvelles pour tous les fournisseurs de services dits « *intermédiaires* »[4] en matière de diligence, de transparence, de coopération avec les autorités publiques, la société civile et les utilisateurs, et de modération des contenus illicites.

Il deviendra applicable à l'ensemble des services concernés à partir du 17 février 2024, et dès le 25 août 2023 aux fournisseurs de très grandes plateformes en ligne et de très grands moteurs de recherche en ligne (*Very large online platforms and search engines*, ou VLOPSEs)[5], dont certains ont été désignés en tant que tels par la Commission européenne le 25 avril 2023[6].

Dans cette phase de transition d'un cadre de régulation national à un cadre européen, l'Arcom a souhaité exercer les prérogatives qu'elle tire de la loi du 24 août 2021 dans le souci d'anticiper la mise en œuvre du RSN. Ainsi, elle s'est fortement inspirée de ce dernier pour la rédaction de ses lignes directrices prises en novembre 2022, dont le présent bilan fait état de la mise en œuvre sur les principales plateformes actives en France. Les analyses réalisées par l'Arcom dans ce bilan des procédures et moyens déployés par les plateformes se fondent sur des observations et sur les rapports que lui ont adressés les opérateurs en réponse à un questionnaire *ad hoc*.

L'Arcom, que le projet de loi visant à sécuriser et réguler l'espace numérique adopté en première lecture par le Sénat le 5 juillet 2023 prévoit de désigner comme coordinateur pour les services numériques (CSN) au titre du RSN pour la France, aux côtés de la CNIL et de la DGCCRF qui participeront à la mise en œuvre de dispositions spécifiques du RSN, entend s'appuyer sur cette première expérience pour contribuer à la mise en œuvre d'une régulation renforcée des plateformes en ligne, en particulier des réseaux sociaux, et au dialogue sur la mise en œuvre du RSN engagé avec les régulateurs européens, i.e. Commission Européenne, coordinateurs pour les services numériques et autres autorités compétentes des États membres.

[4] Les services intermédiaires sont ceux auxquels s'applique le RSN. L'article 3 du texte les définit comme les services de la société de l'information relevant d'une de ces trois catégories :

i) les **services de simple transport,** dont l'activité consiste à transférer les informations à la demande d'un tiers ou à permettre l'accès d'un tiers au réseau (ex. : les fournisseurs d'accès à internet) ;
ii) les **services de mise en cache,** dont l'activité consiste à stocker de manière temporaire des informations pour en faciliter la transmission ultérieure (ex. : les réseaux de diffusion de contenus ou CDN) ;
iii) les **services d'hébergement,** dont l'activité consiste à stocker des données fournies par un tiers, et parmi lesquels on trouve la catégorie spécifique des services de plateforme en ligne (ex : les réseaux sociaux).

[5] Services de plateforme en ligne ou de moteur de recherche en ligne dont l'audience mensuelle est supérieure à 45 millions de destinataires actifs dans l'UE, et désignés par la Commission européenne.

[6] Cette liste, amenée à évoluer à l'avenir, est accessible sur le site de la Commission européenne : https://digital-strategy.ec.europa.eu/en/policies/dsa-vlops

I. Présentation de la démarche de l'Arcom

A. Plateformes concernées

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. Des obligations supplémentaires sont imposées lorsque cette fréquentation dépasse quinze millions de visiteurs uniques[7].

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*. Tous ces services sont des VLOPSEs au sens du RSN, à l'exception de Dailymotion et de *Yahoo Search*, qui ne relèvent pas de ce régime de régulation européenne renforcée.

Le questionnaire faisait notamment référence aux obligations de moyens prévues par le RSN et invitait les opérateurs à faire part d'éventuelles problématiques rencontrées dans la préparation de la mise en œuvre du règlement et dans la mise en œuvre de la loi française.

B. Réponses des opérateurs

Deux opérateurs, *Dailymotion* et *LinkedIn*, ont fait un effort particulier de célérité en transmettant leur contribution bien en amont de la date limite fixée par l'Arcom. En outre, le premier a choisi de fournir une réponse, qui plus est détaillée, malgré une audience qui, selon les données dont il dispose, est inférieure au seuil de déclenchement des obligations de la loi nationale.

De manière générale, l'examen des rapports des plateformes met de nouveau en lumière quatre problématiques déjà rencontrées par l'Arcom[8].

- *Interrogations sur le champ d'application des obligations*

Yahoo a contesté son assujettissement aux dispositions de l'article 42 de la loi du 24 août 2021. D'après ses propres données, l'audience de *Yahoo Search* était inférieure au seuil des 10 millions de visiteurs uniques mensuels en moyenne en France en 2022. En revanche, toujours d'après l'opérateur, l'audience du *Portail Yahoo* (agrégat composé de *Yahoo Homepage*, *Yahoo News*, *Yahoo Style*, *Yahoo Sport* et *Yahoo Finance*) était supérieure à ce seuil sur la période et le territoire considérés. La contribution de l'opérateur concerne donc ce second service. Cependant, l'opérateur considère qu'en raison de son

[7] Seuils définis par le décret n° 2022-32 du 14 janvier 2022 pris pour l'application de l'article 42 de la loi n° 2021-1109 du 24 août 2021 confortant le respect des principes de la République. Conformément à son article 3, « *seules sont prises en compte les connexions à un service, ou à une partie dissociable d'un service, dont l'objet principal est le classement, le référencement ou le partage de contenus mis en ligne par des tiers* ».
[8] En particulier dans le cadre de ses bilans annuels relatifs à la lutte contre la manipulation de l'information sur les plateformes en ligne (disponibles sur le site internet de l'Arcom).

modèle de fonctionnement, le *Portail Yahoo* est un service éditorialisé et non une plateforme en ligne. Dans cette perspective, la loi nationale précitée ne s'appliquerait pas. La déclaration de ce service, par ailleurs peu exploitable, n'a donc pas été prise en compte dans le présent bilan.

- *Prise en compte des différences de modèle des plateformes*

La *Fondation Wikimédia* considère le questionnaire comme inadapté aux particularités du modèle de fonctionnement de *Wikipédia*, collaboratif et décentralisé.

Microsoft explique que *Bing* n'héberge pas les contenus des utilisateurs ni ne permet à ces derniers de publier ou partager du contenu sur le service, et estime de ce fait que certaines questions ne lui sont pas applicables. Cependant, la contribution envoyée par l'opérateur fournit quelques informations intéressantes, notamment en matière de collaboration avec les autorités publiques et d'utilisation du dispositif de signalement par ces dernières.

- *Un effort accru des opérateurs en matière de transparence*

À l'exception de la contribution de *Pinterest*, tous les opérateurs ont accepté la publication de leur réponse dans son intégralité ou quasi intégralité. Cependant, cet effort de transparence se fait parfois au détriment de la pertinence des informations transmises, en particulier dans le cas de *Yahoo*. À l'opposé, *Dailymotion* et *Twitter* ont fourni une contribution substantielle avec un niveau de transparence élevé. De manière générale, **de nombreuses informations ont été communiquées publiquement par les opérateurs**, notamment sur des sujets inédits (ex. : nombre, localisation et langues de travail des modérateurs francophones).

- *Difficultés des opérateurs pour répondre dans les délais impartis*

La grande majorité des plateformes interrogées n'a pas été en mesure de répondre sous deux semaines et a demandé un délai supplémentaire. En particulier, les contributions de *Google* et *TikTok* font état des difficultés pour rassembler et communiquer dans les temps l'ensemble des données demandées, en particulier des données chiffrées – ce qui peut paraître surprenant au vu de leur statut de VLOPSEs et de leurs ressources. L'Arcom estime que la transparence est un élément essentiel du régime de responsabilité des plateformes en ligne en matière de lutte contre les contenus et comportements illicites et qu'à cet égard, les moyens qu'il leur est demandé de déployer pour contribuer à cette lutte doivent aussi porter sur cette transparence et leur capacité à assurer cette dernière avec diligence.

Préconisation :

- ❖ renforcer les moyens adéquats, notamment procéduraux, humains et technologiques, afin de satisfaire aux obligations de transparence avec diligence.

C. Objectifs et méthodologie des observations complémentaires effectuées

En complément des déclarations des plateformes, l'Arcom a procédé à des observations 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 leurs applications.

Ces observations ont porté sur l'accessibilité des conditions générales[9] (CG) des services et la transparence des politiques de modération décrites dans ces dernières, en prenant en compte :

- l'accessibilité des CG depuis toute page du service,
- la mention de l'interdiction de mettre en ligne des contenus à caractère haineux,
- la mention de l'interdiction d'abuser des dispositifs de signalement,
- la description, en des termes clairs et précis, des dispositifs de modération utilisés en la matière (procédures, moyens utilisés, typologie des sanctions, mention des voies de recours, etc.),
- la transparence sur les procédures de suspension et résiliation de comptes ayant mis en ligne, de manière répétée, des contenus à caractère haineux, et le niveau de clarté et de précision de cette présentation.

Les dispositifs de signalement mis à disposition du public sur les services ont également été examinés, en particulier :

- leur accessibilité (y compris par des personnes non connectées à un compte sur le service),
- leur facilité d'utilisation,
- la précision et l'exhaustivité des intitulés de leurs motifs en matière de contenus à caractère haineux.

[9] Ainsi que des règles communautaires et centres de transparence (ou équivalents) des services.

II. Analyse des moyens mis en œuvre par les plateformes en ligne pour lutter contre la haine en ligne

A. Transparence et clarté des conditions générales (CG) sur les règles et les conditions d'application de la modération

Sur les plateformes en ligne, les utilisateurs sont tenus de respecter les lois nationales applicables, ainsi que les éventuelles règles spécifiques à la plateforme. À défaut, la plateforme peut être conduite à modérer le contenu ou le compte contrevenant. Afin de garantir un environnement « *sûr, prévisible et fiable* »[10], la loi française oblige les opérateurs à énoncer au sein des conditions générales de leur service, de manière claire et facilement accessible, les règles et conditions d'application de celles-ci.

L'article 14 du RSN (« *Conditions générales* ») impose à l'ensemble des fournisseurs de services intermédiaires d'énoncer, dans leurs conditions générales, des « *renseignements relatifs aux éventuelles restrictions qu'ils imposent* » aux destinataires du service, informations qui doivent être énoncées « *dans un langage clair, simple, intelligible, aisément abordable et dépourvu d'ambiguïté* » et être mises à disposition du public « *dans un format facilement accessible et lisible par une machine* ». L'obligation de clarté est renforcée lorsque le service s'adresse principalement à des mineurs. En outre, les VLOPSEs doivent notamment (i) publier leurs CG « *dans les langues officielles de tous les États membres dans lesquels elles proposent leurs services* » et (ii) fournir aux destinataires de leurs services un « *résumé des conditions générales* ».

Suivant la même logique, la loi française du 25 août 2021 prévoit que les « *conditions générales d'utilisation du service* » soient facilement accessibles au public et mentionnent :

- l'interdiction de mettre en ligne des contenus haineux illicites[11] ;
- les dispositifs de modération de ces contenus, « *en des termes clairs et précis* » ;
- les mesures à l'égard des utilisateurs qui ont mis en ligne ces contenus et les recours internes et judiciaires dont ils disposent ;
- « *en des termes clairs et précis* », les procédures de suspension ou de résiliation de comptes des utilisateurs ayant mis en ligne, de manière répétée, des contenus illicites (si la plateforme met en œuvre de telles procédures).

Parmi les CG des services observés, seules celles de *Pinterest* mentionnent explicitement la loi française, dans une section dédiée en bas de page qui renvoie le lecteur vers le formulaire de signalement dédié. La loi nationale est également mentionnée dans les « Conditions d'utilisation supplémentaires de la recherche *Google*[12] ». De même, un lien de redirection vers un article du centre d'aide consacré à ce texte de loi est proposé en bas de page des CG d'*Instagram*.

[10] Selon les termes de l'article 1er du RSN.
[11] Dont le champ est clairement défini par le 7° du I de l'article 6 de la LCEN.
[12] NB : celles-ci ne sont valables que pour la France.

1. *L'accessibilité des CG*

Trois pratiques se distinguent quant à l'intitulé de la section permettant à l'utilisateur d'accéder aux CG du service : la majorité des services observés (*Dailymotion*, *YouTube*, *LinkedIn*, *Facebook*, *Pinterest*, *Snapchat*, *Wikipédia* et *Twitter*) a fait le choix d'une mention explicite, et logique, des termes « Conditions générales » ou « Conditions d'utilisation » ; *Google Search*, *Instagram* et *TikTok* ont, pour leur part, privilégié le terme plus général de « Conditions » alors que *Bing* a fait le choix moins explicite du terme « Légal ».

Au vu des observations menées par les services de l'Arcom, le degré d'accessibilité des CG des services apparaît satisfaisant via un navigateur web sur ordinateur : elles sont accessibles en un (*Twitter*, *Instagram*, *Google Search*, *YouTube*, *Dailymotion*, *Wikipédia, TikTok*, et *Bing*) ou deux clics (*Facebook*, *LinkedIn* et *Pinterest*) depuis la page d'accueil du service. Deux clics au maximum sont également généralement nécessaires lorsque l'utilisateur n'est pas connecté à un compte sur le service. Toutefois, la visibilité de la section permettant d'accéder aux CG du service sur cette page d'accueil, comme sur les autres pages du service, pourrait être améliorée. Ce constat est particulièrement vrai pour *YouTube*, et ce que l'utilisateur soit connecté au service ou non.

L'accès aux CG est souvent légèrement plus complexe sur mobile via un navigateur web, avec un nombre maximal de 4 clics requis (*Facebook*, *Instagram* et *Twitter*).

Les CG sont encore moins aisément consultables sur les applications de certains services alors que ces applications sont le premier mode d'accès à ces services : 3 clics sont nécessaires depuis la page d'accueil sur les applications *Snapchat, LinkedIn* et *Facebook*, 4 sur *Twitter, Bing* et *Pinterest* et 5 sur les applications *Dailymotion*, *Instagram* et *TikTok*. De plus, le parcours de navigation est souvent loin d'être intuitif. Un utilisateur devra ainsi faire défiler un long menu déroulant sur *Snapchat*, *Facebook* et *TikTok*. Sur *Dailymotion*, il devra se rendre dans la section « Réglages ». Il en est de même sur *Twitter*, où il lui faut ensuite cliquer sur « Ressources supplémentaires ». Sur *LinkedIn*, il devra se rendre dans la section « Préférences », sur *Instagram* dans la section « À propos ». Enfin, sur *Pinterest*, il devra tout d'abord cliquer sur la section « Enregistrées ». Afin de favoriser l'appropriation des règles en vigueur sur ces services par leurs utilisateurs, les intitulés pourraient être plus explicites et le parcours permettant d'accéder aux CG fluidifié.

Préconisation :

- rendre plus explicite, plus rapide et plus fluide le parcours permettant aux utilisateurs d'accéder aux conditions générales du service.

2. *L'intelligibilité des CG*

Au regard de la nature même des contenus à caractère haineux, proposer des CG intelligibles est à la fois un défi et une nécessité.

Certains opérateurs souffrent de la comparaison en termes de longueur de certaines phrases (*Twitter* et *Snap*), de manque de fluidité en raison de la traduction en français (*LinkedIn*) ou de rareté des exemples fournis (*Pinterest*).

Dans la grande majorité des cas, les CG sont disponibles en français, à l'exception de celles de *LinkedIn* où certaines sections sont uniquement accessibles en anglais.

Une bonne pratique de certains services consiste à fournir des synthèses : sur *LinkedIn*, chaque début de section comprend un encart et une icône en forme d'ampoule, résumant brièvement et simplement le contenu de la section. Sur *Pinterest* et *TikTok*, à la fin de chaque section est proposé un paragraphe en résumant le contenu. Dans le cas de *TikTok*, la visibilité de ces paragraphes pourrait toutefois être améliorée. Ces opérateurs semblent avoir ainsi anticipé l'une des obligations imposées par l'article 14 du RSN aux très grandes plateformes en ligne, qui prévoit que les VLOPSEs fournissent aux destinataires de leur service un « *résumé des conditions générales* ».

Pour plusieurs services, la mise en page des CG pourrait être significativement améliorée afin d'en favoriser l'intelligibilité. Ainsi, la page des CG de *Snapchat* est particulièrement longue et ne propose aucun dispositif de navigation en son sein (renvoi entre sections) ; les conditions applicables aux utilisateurs résidant au sein de l'Union européenne apparaissent après celles applicables aux résidents des États-Unis, alors qu'une présentation géolocalisée des conditions pertinentes éviterait toute confusion. Pour *Bing*, la page énonce les CG de l'ensemble des très nombreux services opérés par *Microsoft* et ne propose aucun renvoi. Enfin, la mise en page des CG des services de *Meta* ne favorise pas leur intelligibilité (longueur, format non « justifié », aucun renvoi vers les sections, etc.).

Les CG de *TikTok*, service massivement utilisé par les mineurs, proposent une mise en page aérée, une syntaxe relativement courte et un vocabulaire peu technique, ce qui est en phase avec l'exigence d'intelligibilité renforcée posée par le RSN. Les résumés du contenu de chaque section en des termes clairs et aisément compréhensibles au sein des CG de *TikTok* et *Pinterest* participent également de cette logique.

Préconisation :

- ❖ améliorer la lisibilité et les conditions de navigation au sein et entre les différentes pages consacrées aux conditions générales, aux règles communautaires, au centre d'aide (ou équivalents), etc.

3. Renseignements relatifs aux restrictions

i) Mention de l'interdiction de publier des contenus à caractère haineux

Deux façons de procéder des plateformes peuvent être distinguées :

- mention claire et explicite de l'interdiction de publier des contenus haineux dans les CG (*TikTok*, *YouTube*, *Dailymotion* et *Microsoft*) ;
- mention concise et/ou générale (ex. : interdiction de publier des contenus illicites) dans les CG avec renvoi vers d'autres pages (ex. : règles communautaires du service) où l'interdiction de publier des contenus haineux est précisée (*Twitter*, *Facebook*, *Instagram*, *Snapchat*, *LinkedIn*, *Wikipédia*[13], *Pinterest* et *Google Search*).

[13] Le lecteur est renvoyé vers le « *Code universel de bonne conduite* », exclusivement disponible en anglais.

À l'exception de *Dailymotion*, *YouTube* et des « Conditions d'utilisation supplémentaires de la recherche *Google* », les CG des opérateurs utilisent une mention générale (ex. : « contenus illégaux ») et/ou limitative (ex. : « harcèlement ») pour désigner les contenus à caractère haineux et n'explicitent pas l'étendue du champ couvert par cette notion. Or, il convient de rappeler que le niveau de précision des conditions générales du service, en des termes accessibles à tous, conditionne la bonne compréhension des règles en vigueur par ses destinataires.

Préconisation :

- préciser clairement, au sein des conditions générales du service, ce que recouvrent les contenus et comportements proscrits par le droit national et les règles de l'opérateur, et notamment l'interdiction des incitations à la haine et des comportements de harcèlement en ligne.

ii) Description des politiques et des dispositifs de modération des contenus à caractère haineux

Les CG des services observés ne font pas état d'éventuelles politiques de modération propres aux contenus à caractère haineux mais traitent en général de la modération des contenus préjudiciables et/ou illicites. Plusieurs pratiques se distinguent en la matière.

Les CG de *Pinterest*, *LinkedIn* et *Bing* se contentent d'explications minimales, de l'ordre de quelques lignes, présentant en des termes clairs les grandes lignes de leur démarche (ex. : énumération de certains contenus pouvant faire l'objet d'une décision de modération et mention de quelques sanctions) et invitent le lecteur à consulter des pages annexes (ex. : centre d'aide ou règles communautaires) pour approfondir le sujet.

Si elles sont davantage détaillées, celles de *Twitter* pourraient être précisées et illustrées. À titre d'exemple, un des motifs peu compréhensibles de modération est rédigé en ces termes : « *notre prestation de Services à votre intention n'est plus commercialement viable* »[14].

Les CG de *Snap*, *Meta*, *TikTok* et *Google* présentent de façon plus complète et précise la démarche pouvant conduire l'opérateur à suspendre ou supprimer un contenu ou un compte. Y sont ainsi évoqués, de manière plus ou moins brève, les dispositifs, mesures et procédures en vigueur sur la plateforme, parfois à l'aide de renvois vers des pages annexes précisant les politiques de la plateforme.

Dans le cas de *Snap*, la description des moyens de modération employés, notamment pour détecter et examiner un contenu, reste évasive et gagnerait à être détaillée. À l'inverse, les CG des services de *Meta*, *Google* et *TikTok* mentionnent explicitement le recours à des outils automatisés à des fins de modération. *TikTok* précise d'ailleurs que sa démarche repose sur une combinaison de ces outils automatisés, de la modération humaine et des signalements adressés par les utilisateurs.

[14] « 4. Utilisation des Services », section « Résiliation des présentes Conditions ». Source : https://twitter.com/fr/tos

En raison de l'approche collaborative et décentralisée de sa politique de modération, comparer *Wikipédia* à d'autres services apparaît peu pertinent. Toutefois, la description de cette politique dans les CG de la plateforme mériterait de gagner en précision et en transparence.

Dailymotion fait preuve d'un degré de transparence significatif. Une annexe à ses CG intitulée « Politique relative aux Contenus Prohibés » détaille ainsi successivement (i) les différentes catégories de contenus prohibés (ex. : « D. Contenus haineux »), (ii) la détection et le signalement de ces contenus (dispositifs de détection automatique, outil de signalement dédié, signalement par e-mail ou voie postale) et (iii) les conséquences en cas de non-respect de la politique relative aux contenus prohibés. Cette dernière partie présente tout d'abord les actions de modération applicables aux contenus prohibés (affectant la disponibilité, la visibilité, l'accessibilité ou la monétisation de ceux-ci) avant d'expliciter le processus d'appel d'une décision de modération. Ce haut niveau de transparence et cet effort didactique semblent à même d'assurer une bonne compréhension des règles en vigueur par les utilisateurs.

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 (*Twitter*, *TikTok*, *Facebook*, *Instagram*, *Google Search*, *YouTube*, *Dailymotion*, *Fondation Wikimédia*) ou dans les règles communautaires (*LinkedIn*). D'après ces CG, les réclamations peuvent porter soit sur des décisions de modération de comptes (*Twitter* et *Google Search*), soit sur des décisions de modération de contenus (*Wikipédia*, *Dailymotion* et *Instagram*), soit sur les deux types de décisions (*TikTok*, *YouTube*, *Facebook* et *LinkedIn* via ses « Politiques de la communauté professionnelle »).

Parmi ces services, et à l'exception de *Wikipédia* et de *Dailymotion*, les CG réservent cette faculté de réclamation aux décisions « positives » de modération (ex. : suspension ou suppression d'un compte ou contenu) en excluant les décisions de ne pas donner suite à un signalement.

En revanche, trois plateformes font figure d'exceptions :

- *Pinterest* : la possibilité de faire appel d'une décision de modération n'est pas clairement mentionnée dans les CG. Il faut atteindre le centre d'aide pour trouver des informations à ce sujet ;
- *Snapchat* et *Bing* : aucune mention de la possibilité de faire appel d'une décision de modération ne figure dans les CG ni, concernant *Snapchat,* dans les règles communautaires.

Cette pratique ne semble pas conforme aux dispositions de l'article 14 du RSN qui précisent que les CG devront notamment présenter des informations sur « *le règlement intérieur de leur système interne de traitement des réclamations* ».

En outre, les sections dédiées au sein des CG de *Google Search*, *YouTube*, *Facebook*, *Instagram*, *Twitter* et *LinkedIn* (via ses « Politiques de la communauté professionnelle ») renvoient le lecteur vers une page d'aide relative aux voies de recours. Il en va de même de *Wikipédia*, mais la page concernée est uniquement disponible en anglais. Enfin, au sein des CG de *TikTok*, la section relative aux « Droits de TikTok », qui évoque ce sujet, comporte un lien vers un formulaire de contact permettant à toute personne de faire appel d'une décision de modération.

De manière générale, les systèmes internes de traitement des réclamations d'une décision de modération sont accessibles (i) après une navigation complexe au sein des CG et/ou pages d'aide du service, (ii) via une requête internet dédiée (ex. : « appel suspension/suppression compte/contenu [nom du service] ») ou (iii) via l'envoi d'un message à l'utilisateur ayant fait l'objet de la décision de modération[15].

Enfin, les CG de *Dailymotion* et *YouTube* et les « Conditions d'utilisation supplémentaires de la recherche *Google* » mentionnent explicitement les voies de recours judiciaires, ce qui paraît de nature à garantir l'information la plus complète de l'utilisateur.

En outre, les CG de certains services mentionnent explicitement le droit dont disposent leurs opérateurs de notifier un contenu illicite à un tiers, y compris aux forces de l'ordre (*Snapchat*, *Facebook* et *Instagram*) et aux autorités judiciaires compétentes (*Dailymotion* et *Wikipédia*).

Préconisation :

- être le plus clair et explicite possible, au sein des conditions générales, sur l'existence et le fonctionnement du mécanisme interne de recours des décisions de modération tant pour les personnes qui signalent un contenu que pour celles dont le contenu est l'objet d'une modération.

B. Dispositifs de signalement des contenus à caractère haineux

Les opérateurs de plateforme en ligne ne sont pas soumis à une obligation générale de surveillance des contenus publiés sur leur service mais sont tenus de retirer ceux d'entre eux qui seraient manifestement illicites dès lors qu'ils leur ont été signalés. Afin de permettre une modération efficace, la loi oblige notamment les opérateurs à mettre à disposition un outil de signalement conçu de telle façon qu'il facilite au maximum l'acte de signalement.

L'article 16 du RSN prévoit que les « *fournisseurs de services d'hébergement mettent en place des mécanismes permettant à tout particulier ou à toute entité de leur signaler la présence au sein de leur service d'éléments d'information spécifiques que le particulier ou l'entité considère comme du contenu illicite* ». Ces mécanismes doivent être « *faciles d'accès et d'utilisation* » et concourir à faciliter la soumission de notifications « *suffisamment précises et dûment étayées* ».

Poursuivant le même objectif, la loi française du 24 août 2021 demande aux plateformes de mettre en place « *un dispositif, aisément accessible et facile d'utilisation, permettant à toute personne* » de porter à leur connaissance, « *par voie électronique* », un contenu considéré comme présentant un caractère haineux.

Deux services se singularisent en la matière.

Wikipédia, en raison des spécificités de son modèle, ne dispose pas de dispositif de signalement « traditionnel ». Confronté à un contenu problématique, un utilisateur pourra (i) le corriger lui-même en modifiant la page, (ii) expliquer le problème sur la page de

[15] Pour une analyse des voies de recours à disposition d'un destinataire du service, voire le II. D).

discussion de l'article concerné ou (iii) demander de l'aide sur le « Forum des nouveaux »[16]. L'architecture de la modération sur la plateforme est remarquable, en ce qu'elle repose en priorité sur les utilisateurs, tout en assurant une expertise stable en matière de modération des contenus illicites grâce aux utilisateurs « administrateurs » de *Wikipédia*[17].

Sur *Bing*, un utilisateur souhaitant signaler un contenu illicite est redirigé vers le formulaire proposé par l'association Point de Contact sur son propre site internet[18]. Ce dispositif a pour inconvénient de ne pas permettre de saisir directement les équipes de modération de l'opérateur de la plateforme. Certes, un formulaire de signalement est proposé par *Microsoft* mais il est peu aisément accessible[19] et, surtout, il ne comporte aucun motif permettant de signaler un contenu à caractère haineux.

Par ailleurs, parmi les services observés, seules les CG de *Wikipédia*, *Dailymotion*, *YouTube*, *Facebook*, *Instagram* et les « Conditions d'utilisation supplémentaires de la recherche *Google* » mentionnent explicitement l'interdiction d'abuser de cet outil de signalement.

1. *Les formulaires de signalement sont-ils faciles d'accès pour l'utilisateur ?*

i) Pour un utilisateur non connecté à un compte sur le service

La loi française prévoit que le dispositif de signalement du service soit accessible « *à toute personne* ». Cela inclut donc les utilisateurs non connectés à un compte, lorsque le service est accessible hors inscription et connexion. Il convient de noter que cette obligation soulève un enjeu technique spécifique pour les opérateurs : l'émetteur du signalement n'étant pas connecté à la plateforme, l'opérateur ne connaît pas ses coordonnées et doit donc déployer un dispositif spécifique pour recueillir ces signalements dans les conditions exigées par la loi (c'est-à-dire, en accusant réception à l'utilisateur et en l'informant des suites données au signalement).

Google Search, *YouTube* et *TikTok* proposent un formulaire *ad hoc* permettant à toute personne de signaler un contenu à caractère haineux au titre l'article 6-4 de la LCEN[20]. L'Arcom souligne l'intérêt de cette pratique, mais estime que l'accessibilité de ces formulaires pourrait être améliorée dans la mesure où (i) de nombreux clics sont nécessaires pour accéder à celui de *Google Search*, (ii) celui de *YouTube* n'est accessible qu'après une navigation peu aisée au sein des pages d'aide et (iii) celui de *TikTok* présente des difficultés de chargement empêchant son affichage.

À cet égard, l'Arcom estime que l'existence de formulaires spécialement dédiés à l'article 6-4 de la LCEN ne devrait pas conduire la plateforme à considérer, dans ses rapports de transparence, que seuls les contenus signalés par le biais de ces formulaires sont à comptabiliser comme contenus à caractère haineux signalés sur le service en France. En effet, l'existence de ces formulaires n'est pas connue de tous, et leur intitulé peut décourager les utilisateurs. Il convient donc de considérer également ceux signalés au titre

[16] Source : https://fr.wikipedia.org/wiki/Aide:Accueil/Signaler
[17] Source : https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Administrateur
[18] Source : https://www.pointdecontact.net/cliquez-signalez/
[19] Uniquement accessible via une requête internet dédiée (« signalement Bing ») et plusieurs clics. En revanche, la requête « formulaire signalement Bing » ne permet pas, et ce de manière contrintuitive, d'y accéder.
[20] Issu de la loi du 24 août 2021 - voir note n° 5.

des règles en vigueur sur la plateforme via le dispositif de signalement à proximité directe des contenus.

Les services opérés par *Meta* disposent également d'un formulaire dédié aux signalements au titre de la LCEN. Cependant (i) il n'est accessible qu'après une navigation complexe et peu explicite au sein du dispositif de signalement intégré puis des pages d'aide du service ou une requête internet très précise (ex. : « formulaire signalement respect république Facebook ») et (ii) il est partiellement rédigé en anglais. Ainsi, son accessibilité au public pourrait être améliorée.

Modifié courant 2022, le dispositif de signalement sur *Dailymotion* permet dorénavant aux utilisateurs non connectés de signaler un contenu. *Snapchat*, pour sa part, semble disposer d'un formulaire *ad hoc* permettant à toute personne de signaler un « abus ». Cependant, lors d'une observation réalisée en juin 2023, la page était inaccessible.

Enfin, *Twitter*, *Pinterest* (sauf via un formulaire *ad hoc*) et *LinkedIn* ne proposent pas de dispositif permettant aux utilisateurs non connectés de signaler des contenus illicites.

Préconisation :

- dans les rapports de transparence, comptabiliser à la fois :
  - les contenus à caractère haineux signalés par le biais de formulaires *ad hoc* et via le dispositif de signalement à proximité directe des contenus ;
  - ceux retirés (i) sur la base des conditions générales et (ii) au titre du droit national.

ii) Pour un utilisateur connecté à un compte sur le service

Tous les services observés possèdent des dispositifs de signalement mis à disposition des utilisateurs connectés à un compte. Ces outils sont accessibles en moins de deux clics depuis tout contenu du service[21]. Ils permettent généralement de finaliser le signalement d'un contenu à caractère haineux en moins de cinq clics, à l'exception de *Twitter* et *Google Search* pour lesquels davantage de clics sont nécessaires.

Cependant, sur de nombreux services (*YouTube*, *LinkedIn*, *Facebook*, *Instagram*, *Pinterest*, *Twitter*, *TikTok*), l'accès au dispositif de signalement est conditionné au clic sur un bouton dont l'intitulé est peu explicite (ex. : « … », voir capture d'écran ci-dessous).

[21] À quelques exceptions près ; par exemple, il n'est pas possible de signaler un « Événement » sur Facebook.

Capture d'écran de l'application Twitter, le 18 juillet 2023 (source : Arcom)

L'accessibilité de ces dispositifs devrait donc être améliorée. *Snapchat* propose également des intitulés parfois peu explicites : à titre d'exemple, un utilisateur souhaitant signaler un profil « ami » en raison d'un comportement illicite devra préalablement cliquer sur l'intitulé « Gérer l'amitié ». À l'inverse, le bouton permettant d'accéder au dispositif de signalement d'un contenu est bien visible et explicite sur *Dailymotion*.

Pinterest dispose d'un formulaire de signalement dédié permettant à tout utilisateur connecté de signaler un contenu au titre de la LCEN. Toutefois, l'intitulé du motif permettant d'y accéder, « Violation de la loi confortant le respect des principes de la République », est très peu parlant pour le public. Dans un souci d'intelligibilité, il serait souhaitable d'expliciter davantage cet intitulé.

iii) Cas particulier du signalement d'un compte sur les plateformes de partage de vidéo

L'accessibilité du dispositif de signalement des comptes sur *YouTube* pourrait être grandement facilitée. En effet, sur la page d'un compte, l'idée d'aller dans la dernière section, intitulée « À propos », ne semble pas intuitive. De plus, l'accès au dispositif de signalement n'est alors matérialisé que par une icône en forme de drapeau de taille modeste et en noir et blanc.

Sur *Dailymotion*, il n'est pas possible de signaler un compte.

Préconisation :

- ❖ améliorer l'accessibilité des dispositifs de signalement, notamment par des symboles et des intitulés plus explicites.

2. *L'intitulé des motifs de signalement est-il clair ?*

Le degré de clarté de l'intitulé des motifs de signalement d'un contenu à caractère haineux varie d'un service à l'autre.

Il est élevé lorsque le service possède un formulaire de signalement et/ou une catégorie de signalement dédié aux contenus à caractère haineux au titre de la LCEN. C'est le cas de *Google Search*, *YouTube*, *Pinterest*, *Facebook*, *Instagram* et *TikTok*. En outre, *Pinterest*, *Google Search* et *YouTube* se distinguent par des intitulés concis alors que les intitulés, plus longs, du formulaire dédié accessible via les services de *Meta* indiquent la disposition légale à laquelle ils correspondent. Par ailleurs, deux motifs des formulaires de *Google Search* et *YouTube* sont illustrés par des exemples. De plus, le premier renvoie l'utilisateur vers le texte de loi et le second, au moment d'indiquer le motif du signalement, vers une page consacrée au signalement de contenus à caractère haineux au titre de la LCEN.

Quant aux dispositifs de signalement des contenus à caractère haineux intégrés aux services de *Dailymotion*, *Meta*, *YouTube* et *LinkedIn*, ils sont construits avec clarté. Les motifs proposés sont par ailleurs illustrés par des exemples. Celui de *Snapchat* est également aisément compréhensible : l'énumération des motifs de signalement est longue, mais ces derniers sont classés en sous-ensembles afin d'en favoriser l'ergonomie.

Le périmètre des contenus haineux au titre de la loi française englobant des infractions de différents ordres, les plateformes font le choix, par souci de lisibilité pour l'utilisateur, de couvrir certaines de ces infractions par d'autres intitulés.

S'agissant de *Twitter*, en revanche, la distinction entre les intitulés des différents motifs de signalement d'un contenu n'est pas toujours intuitive. À titre d'exemple, un contenu à caractère haineux semble pouvoir correspondre simultanément aux deux sous-motifs suivants : « *L'utilisateur menace de faire usage de la violence ou de blesser quelqu'un* » et « *Les propos incitent à la haine envers une catégorie protégée* ». Cette ambiguïté peut conduire à décourager l'utilisateur de procéder au signalement.

Pour ce qui est de *Google Search*, il convient tout d'abord de noter que la présentation de l'étape suivant la sélection du produit Google auquel le signalement se rapporte est confuse : l'utilisateur est amené à cocher « *Oui* » ou « *Non* » sans qu'aucune question ne lui soit posée. Or la réponse aura une incidence sur la suite du processus en permettant d'accéder, ou non, au formulaire consacré aux « *motifs non juridiques liés au règlement pour signaler du contenu* ». En outre, dans ce dernier formulaire, le motif « *Doxxing : signalez du contenu présentant vos coordonnées et contenant des menaces explicites ou implicites, ou des incitations à l'action explicites ou implicites visant à nuire ou à harceler* » pourrait être rendu plus explicite.

Enfin, la majorité des opérateurs (*Dailymotion*, *LinkedIn*, *Meta*, *Microsoft*, *Pinterest*, *Snap*, *TikTok* et *Twitter*) n'a pas fait état de difficultés particulières pour inclure les motifs liés à la haine en ligne à leurs dispositifs de signalement. En revanche, *Google* a rappelé la nécessité de trouver un équilibre entre simplicité d'utilisation du système de signalement et mise en place de solutions conformes d'un point de vue juridique.

Cependant, il apparaît que la précision des intitulés des motifs de signalement pourrait parfois être améliorée afin de couvrir plus explicitement le périmètre de la loi nationale. À titre d'exemple, si tous les services disposent d'un motif couvrant le harcèlement en

général, aucun ne précise explicitement qu'il peut s'agir de harcèlement scolaire[22]. On peut néanmoins relever que *Dailymotion* précise que le motif « *Maltraitance des enfants* » inclut le harcèlement en ligne, et que *Facebook* se distingue par un message apparaissant lors d'un signalement pour harcèlement qui mentionne spécifiquement la protection renforcée offerte par ses règles à l'égard des mineurs. À l'inverse, *Twitter* possède un dispositif de signalement sur lequel l'utilisateur devra nécessairement cliquer sur un premier intitulé (« *Les propos tenus sont inappropriés ou dangereux* ») pour accéder au motif dédié au harcèlement.

3. *Quelques bonnes pratiques*

i) Sur l'ensemble des services observés

Une bonne pratique fréquemment observée réside dans l'explicitation et l'illustration des intitulés du dispositif de signalement (ex. : *Dailymotion*, *Google Search*, *LinkedIn*, *Facebook*, *Instagram* et *Pinterest*) et dans le rappel succinct des règles en vigueur en matière de contenus (ex. : *Google Search*, *Facebook*, *Instagram*, *Pinterest*, *Snapchat* et *TikTok*) avant que l'utilisateur ne finalise son signalement.

Peut être considérée comme une bonne pratique la possibilité laissée à l'utilisateur d'adjoindre à son signalement une description de tout élément qu'il jugera utile lors de l'analyse de celui-ci (*Google*, *YouTube*, *Pinterest* et *Twitter*). À l'inverse, on peut s'interroger sur le caractère potentiellement dissuasif de la pratique constatée sur *Dailymotion* et *Snapchat* consistant à rendre obligatoire une telle description, même sommaire.

Les formulaires de signalement dédiés à l'article 6-4 de la LCEN (*Google Search*, *YouTube*, *Pinterest*, *Facebook*, *Instagram* et *TikTok*) proposent également cette fonctionnalité, de manière facultative à l'exception de *YouTube*, anticipant ainsi l'une des dispositions de l'article 16 du RSN qui prévoit que les notifications émises par le biais du dispositif de signalement contiennent « *une explication suffisamment étayée des raisons pour lesquelles le particulier ou l'entité allègue que les informations en question sont du contenu illicite* ».

La possibilité offerte par *Pinterest* aux utilisateurs signalant un contenu par le biais du formulaire dédié à l'article 6-4 de la LCEN d'indiquer la ou les partie(s) du contenu que leur signalement concerne (ex. : « *Image de profil* », « *Nom de profil* », « *Description de profil* » ou « *Autre…* ») est de nature à améliorer la précision des informations transmises.

La fonctionnalité proposée par *LinkedIn* permettant à l'utilisateur d'indiquer s'il souhaite recevoir des nouvelles sur le statut de son signalement constitue également une bonne pratique que l'Arcom avait d'ailleurs recommandée dans ses lignes directrices.

Enfin, en permettant à un utilisateur de signaler plusieurs tweets d'un même compte via un seul et unique signalement, *Twitter* se distingue par une pratique qui semble à même de fluidifier le parcours de signalement.

[22] Le harcèlement scolaire constitue une infraction au titre du Code pénal (article 222-33-2-3) depuis la loi n° 2022-299 du 2 mars 2022 visant à combattre le harcèlement scolaire, qui l'a par ailleurs inclus parmi les motifs de haine en ligne au sens de la LCEN.

ii) Cas particulier des plateformes de partage de contenus vidéo

Dailymotion permet à tout utilisateur effectuant un signalement, qu'il soit connecté ou non au service, d'indiquer l'horodatage à partir duquel il conviendrait, selon lui, de concentrer l'analyse du contenu vidéo. *YouTube* offre la même fonctionnalité. Cependant, sur cette plateforme, le signalement ne peut être effectué que par un utilisateur connecté au service. Cette fonctionnalité gagnerait encore en pertinence si elle permettait à l'utilisateur émettant un signalement d'indiquer un intervalle de temps complet.

Enfin, une autre bonne pratique réside dans la possibilité offerte par *YouTube* à ses utilisateurs connectés d'indiquer si leur signalement s'applique également aux liens inclus dans la description de la vidéo signalée.

Préconisations :

- illustrer les intitulés des motifs de signalement par des exemples concrets ;
- rappeler, succinctement, les règles en vigueur sur le service en matière de contenus avant que l'utilisateur ne finalise son signalement ;
- offrir à l'utilisateur la possibilité de joindre à son signalement une description de tout élément qu'il jugera utile à l'analyse de celui-ci ;
- offrir à l'utilisateur la possibilité d'indiquer la ou les partie(s) du contenu que son signalement concerne ;
- permettre à l'utilisateur de signaler plusieurs contenus d'un même compte via un seul et unique signalement ;
- permettre à l'utilisateur signalant un contenu vidéo, d'indiquer un intervalle de temps complet et pas uniquement un horodatage de début ;
- offrir à l'utilisateur la possibilité d'indiquer si son signalement s'applique également aux liens inclus dans la description du contenu signalé.

En outre, conformément à ce qu'elle avait déjà indiqué dans le cadre de ses lignes directrices, l'Arcom rappelle la préconisation suivante :

- permettre à l'utilisateur d'indiquer s'il souhaite être informé de l'évolution du traitement de son signalement.

4. Lutte contre les signalements abusifs

L'article 23 du RSN imposera aux fournisseurs de services d'hébergement de prendre des mesures pour lutter contre l'usage abusif des dispositifs de notification de contenus illicites. L'article 6-4 de la LCEN prévoit pour sa part que les plateformes peuvent suspendre temporairement ou, dans les cas les plus graves, définitivement, les utilisateurs pratiquant des signalements abusifs.

Ces dispositions visent à protéger les utilisateurs de pratiques frauduleuses, prenant la forme de « *raids* » de signalements massifs dans l'objectif unique de tromper la modération des opérateurs par une suspension injustifiée du compte ciblé.

Toutefois, les opérateurs interrogés signalent avoir pris peu de mesures préventives ou coercitives pour décourager les usages abusifs de leurs outils de signalement. *LinkedIn* ou *Snapchat*, par exemple, affirment présumer la bonne foi de tout signalement d'un contenu.

Twitter et *Meta* précisent être en capacité de suspendre des utilisateurs qui feraient un usage abusif des outils de signalement, mais indiquent avoir fait le choix de ne suspendre aucun compte d'utilisateur pour ce motif en 2022.

Préconisation :

- ❖ être vigilant et, le cas échéant, maintenir une capacité de réaction face à des usages abusifs des outils de signalement.

C. Moyens mis en œuvre pour la modération des contenus à caractère haineux par les plateformes en ligne

Les opérateurs doivent traiter les signalements qu'ils reçoivent avec diligence, en mettant en œuvre les moyens appropriés, notamment en termes de modérateurs. Ils sont également tenus de porter une attention particulière aux notifications provenant des « *signaleurs de confiance* », entités reconnues pour leur expertise en matière de lutte contre les contenus illicites en ligne.

L'article 16 du RSN imposera aux fournisseurs de services d'hébergement, notamment ceux offrant des services de plateformes en ligne, de mettre en place des mécanismes de notification et d'action permettant à tout particulier ou à toute entité de signaler un contenu illicite au sein de leur service. Ces fournisseurs devront garantir le traitement de ces notifications et des contenus illicites qu'elles visent « *en temps opportun, de manière diligente, non arbitraire et objective* ».

L'article 22 du règlement renforce cette obligation de diligence à l'égard des notifications de contenus illicites adressées par des entités désignées « *signaleurs de confiance* », qui doivent faire l'objet d'un traitement prioritaire. Les fournisseurs de service d'hébergement sont tenus de mettre en œuvre des « *mesures techniques et organisationnelles* » adaptées pour répondre « *dans les meilleurs délais* » aux notifications adressées par les signaleurs de confiance, qui assument dès lors un rôle particulier dans l'analyse et la lutte contre la dissémination publique de contenus illicites en ligne.

Il renforce également les obligations procédurales visant à garantir le traitement diligent des signalements adressés par les « *tiers de confiance* », reconnus comme tels au regard de leur « *expertise et compétence particulière aux fins de détection, de l'identification, et du signalement des contenus illicites* », dès lors qu'ils représentent des « *intérêts collectifs* » et présentent des « *garanties de diligence et d'objectivité* ».

1. *Moyens humains et procédures mises en œuvre pour traiter les signalements des utilisateurs*

L'Arcom a interrogé les opérateurs sur les moyens humains mis en œuvre pour assurer le traitement des signalements.

L'Autorité constate qu'une partie significative des opérateurs[23] refuse toujours de divulguer publiquement le nombre de personnes dédiées à cette tâche.

Toutefois, *Twitter* déclare disposer de 149 modérateurs « *dont des francophones* »[24], *Dailymotion* estime à « *une trentaine de modérateurs* » les effectifs concernés en 2022, tandis que *LinkedIn* déclare que son équipe interne est composée d'une « *centaine de réviseurs* », dont « *environ 29 francophones* ». La *Fondation Wikimédia* estime qu'une « *cinquantaine* » de collaborateurs bénévoles contribuent particulièrement à la lutte contre les contenus illicites sur leur plateforme, tout en rappelant que les principes particuliers de la modération collaborative de l'encyclopédie en ligne rendent cette donnée peu pertinente.

L'Autorité relève enfin que les opérateurs qui font preuve de transparence sur leur mécanisme de modération précisent soumettre tout signalement d'utilisateur à un double contrôle, d'une part, au regard de leurs conditions générales ou standards d'utilisation (*Meta, Twitter, TikTok*) et, d'autre part, au regard des dispositions légales applicables en France.

Dans l'hypothèse où un contenu signalé ne contrevient pas aux standards de la communauté ou aux politiques internes de l'opérateur mais serait malgré tout illicite au regard du droit français, le contenu en question est retiré pour la France uniquement.

Meta signale que sur les 4 807 contenus illicites qu'il déclare avoir bloqués en France en 2022 sur ses deux plateformes, seuls 19 contenus l'ont été uniquement en raison d'une violation des dispositions de l'article 6-4 de la LCEN. Cet opérateur accorde une claire priorité aux retraits pour violation des standards de la communauté. Si cette méthode d'examen n'est pas contestable, l'Autorité relève qu'elle a pour effet de minimiser, dans les chiffres rendus publics, le nombre de contenus illicites à caractère haineux réellement modérés.

Préconisations :

- ❖ renforcer la transparence des politiques de modération en rendant publics le nombre, la langue de travail et la localisation des modérateurs employés par l'opérateur ;
- ❖ assurer le dimensionnement adéquat des moyens humains dédiés à la modération des contenus illicites.

2. Reconnaissance de tiers de confiance en France

Au vu de leurs réponses, les opérateurs peuvent être classées en deux ensembles distincts : ceux qui ont établi des liens plus ou moins étroits avec un ensemble de tiers de confiance (*Twitter*, *Meta, Google, TikTok, Snapchat*) et ceux qui déclarent ne pas collaborer avec des tiers de confiance en raison de la nature de leur plateforme (*Microsoft* pour *Bing*, *Yahoo*,

[23] *Google*, *Meta*, *Snapchat*, *Pinterest*, *Yahoo* et *TikTok*

[24] Interrogé par l'Arcom, *Twitter* a précisé qu'il s'agissait du nombre de modérateurs pour la zone Europe, compétents pour intervenir sur les violations des règles de la plateforme ou des lois locales. Ce chiffre peut paraître faible au vu de la taille de la plateforme et de la vivacité qui caractérise les échanges qui s'y déroulent, mais il est difficile de l'apprécier sans pouvoir comparer avec les moyens déployés par les autres grandes plateformes en ligne, qui n'ont pas eu la transparence de *Twitter* à ce sujet dans leur réponse à l'Arcom. L'application du RSN, qui exige complétude et transparence de la part des plateformes, permettra de faire un exercice de comparaison solide.

Fondation Wikimédia) ou pour des motifs d'opportunité relevant de leur politique interne (*Pinterest*, *LinkedIn*).

Dailymotion déclare avoir eu des « *liens anciens désormais distendus* » avec des signaleurs de confiance non précisés, mais être volontaire pour restaurer ceux-ci. Cet opérateur déclare avoir mis en place un canal de signalement dédié aux tiers de confiance au cours de l'année 2022.

3. Modalités de collaboration avec les tiers de confiance

Parmi les opérateurs qui déclarent collaborer avec des signaleurs de confiance établis en France, on constate une forte variabilité du nombre de partenaires. Par ailleurs, il n'y pas de corrélation entre le nombre de signaleurs de confiance partenaires déclaré et le taux de signalement de contenus. En outre, certains opérateurs n'ont pas pu quantifier le nombre de signalements transmis par leurs partenaires de confiance.

Ainsi *Meta*, qui déclare 17 « *partenaires de confiance* » en France, ne précise pas le nombre de signalements de contenus illicites de leur part, tandis que *TikTok* (11 « *signaleurs de confiance* » reconnus) déclare avoir reçu 17 signalements de contenus à caractère haineux au sens de la LCEN pour l'année 2022 de la part de ses partenaires. À l'inverse, *Snapchat,* qui déclare collaborer avec 4 « *tiers de confiance* », signale avoir reçu 1 377 signalements (dont 1 375 transmis par e-Enfance).

Dans le même ordre d'idée, *Twitter*, qui reconnaît 5 signaleurs de confiance en France, signale avoir reçu 242 signalements de contenus à caractère haineux en 2022, et avoir retiré le contenu signalé dans 66 % des cas. Ce chiffre est élevé en comparaison des taux de 24 % et 43 % de décisions de retrait prises par l'opérateur à la suite, respectivement, d'un signalement utilisateur et d'un signalement des forces de l'ordre.

Twitter précise disposer d'un canal de signalement sous forme de portail exclusivement dédié aux signaleurs de confiance, ce qui paraît être une bonne pratique de nature à garantir la traçabilité des signalements et la conformité avec les exigences du RSN en matière de priorisation des notifications adressées par les signaleurs de confiance.

À l'inverse, certains opérateurs ne mentionnent pas l'existence d'un dispositif spécifique de signalement pour garantir un traitement adéquat de ces notifications particulières.

C'est le cas de *Meta* et de *Google*, ce dernier précisant qu'il n'est pas en mesure de transmettre de données chiffrées en rapport avec le nombre de signalements transmis par ces derniers.

L'Arcom constate enfin que les opérateurs n'ont pas transmis de données permettant d'objectiver leur célérité dans le traitement des notifications des signaleurs de confiance. Elle souligne qu'il s'agit d'une obligation substantielle du RSN et une garantie de l'efficacité du dispositif de lutte contre la dissémination des contenus illicites.

Préconisations :

- ❖ systématiser, dans les rapports de transparence, la distinction de l'origine du signalement entre utilisateurs, signaleurs de confiance et autorités publiques ;
- ❖ collecter et rendre publiques les données permettant d'objectiver la rapidité de traitement des notifications transmises par les signaleurs de confiance.

D. Voies de recours

Les opérateurs de plateformes en ligne sont tenus de mettre en place un dispositif permettant à toute personne de contester une décision de modération prise par l'opérateur. Ces dispositifs doivent être facile d'accès et utilisables sans difficulté par tout utilisateur.

Le RSN impose à tous les fournisseurs de services d'hébergement, notamment de plateformes en ligne, de motiver de manière claire et précise toute restriction d'utilisation à l'encontre d'un destinataire du service. Ces restrictions, énumérées à l'article 17 du règlement, peuvent notamment consister dans le retrait d'un contenu illicite ou la suspension ou suppression d'un compte.

Ces décisions doivent cependant pouvoir être contestées par les destinataires du service. L'article 20 du règlement dispose que les fournisseurs de services d'hébergement devront mettre en place un système interne de traitement des réclamations permettant de faire appel des décisions, dans des délais raisonnables et en garantissant un traitement non discriminatoire et non arbitraire.

L'article 6-4 de la LCEN renforce les obligations pesant sur les grands opérateurs de plateformes en ligne en matière de motivation des décisions de retrait de contenu ou de suspension de compte. L'existence des voies de recours doit être clairement exposée dans les conditions générales d'utilisation et les décisions doivent être motivées. Les sanctions telles que la suspension ou la suppression d'un compte utilisateur doivent être proportionnées aux infractions et susceptible de recours.

Le taux de recours des décisions initiales varie d'une plateforme à l'autre mais est faible dans l'ensemble. Les opérateurs sont rares à faire la distinction entre les recours formés contre les décisions de contenus et les recours formés contre les suspensions ou suppressions de comptes, alors que l'atteinte à la liberté d'expression des utilisateurs est plus grave dans ce second cas.

Par ailleurs, le taux d'infirmation de cette décision sont généralement assez faibles. *LinkedIn* affirme que, sur 3 064 contenus à caractère haineux supprimés en 2022, seulement 5 ont fait l'objet d'un recours, et aucune des décisions attaquées n'a été infirmée.

Twitter indique que sur les 11 538 recours reçus au titre de la LCEN en 2022, 600 ont fait l'objet d'une action au terme d'un réexamen, soit 5,2 %. La nature de ces actions (restauration du contenu, du compte, ou à l'inverse, suppression d'un contenu préalablement considéré comme licite) n'est pas détaillée.

TikTok se distingue nettement par un nombre de recours élevé (27 770 demandes) et par un taux d'infirmation des décisions particulièrement élevé, à hauteur de 40 %. Cet ordre de grandeur est similaire pour *Dailymotion* (44 %), sur un nombre de décisions initiales très nettement inférieur. Ces taux très élevés de révision des décisions témoignent, s'il le fallait, de l'utilité des procédures de recours mais surtout, ils soulèvent des interrogations sur la pertinence de la modération initiale.

Certaines plateformes estiment que les recours introduits par les utilisateurs sont de faible valeur ajoutée ; *Meta*, qui relève 625 demandes de réexamen en 2022, explique le fait qu'aucune décision n'ait été infirmée par ses services par « *la faible qualité et l'absence de motifs juridiques* » des demandes de réexamen, tout en déplorant la présence d'un certain nombre de *spams.* L'Autorité note cependant qu'il incombe aux fournisseurs de service de veiller à ce que les dispositifs de contestation des décisions de modération soient « *aisément accessibles et faciles d'utilisation* ».

On relève enfin qu'aucun opérateur ne signale de recours judiciaire formé contre ses décisions.

Préconisations :

- permettre à tout utilisateur de contester une décision de modération et assurer un égal traitement de ces recours, sans que soit exigée une argumentation juridique particulière ;
- lorsque le taux d'infirmations, à la suite d'un recours, des décisions ou actions de modération est élevé (comme c'est le cas pour *TikTok* et *Dailymotion*), prendre les mesures adéquates pour évaluer la pertinence de la modération initiale et y remédier le cas échéant.

E. Devoir de coopération avec les autorités judiciaires et administratives nationales

Les opérateurs de plateforme en ligne sont tenus de mettre en place des procédures et des moyens, humains et technologiques, permettant de garantir une réponse rapide aux sollicitations des forces de l'ordre et de l'autorité judiciaire.

Ces demandes peuvent porter sur un contenu illicite au sens du droit français ou sur des données permettant d'identifier un utilisateur du service suspecté d'avoir diffusé un contenu illicite.

Le RSN impose à tous les fournisseurs de services intermédiaires de répondre aux injonctions des autorités judiciaires et administratives nationales compétentes sur la base du droit de l'Union ou du droit national dans les « *meilleurs délais* » et de les informer de la suite donnée à ces injonctions.

Les autorités judiciaires ou administratives peuvent enjoindre le fournisseur de service d'agir contre un contenu illicite ou de transmettre des informations sur un destinataire du service, au titre des articles 9 et 10 du règlement.

En outre, les fournisseurs de services intermédiaires, s'ils ne sont pas assujettis à une obligation générale de surveillance ou de recherche active de faits ou d'activités illicites sur leurs services, sont astreints à une obligation de notification des soupçons d'infraction pénale aux autorités judiciaires ou répressives, dès lors qu'il y a lieu de craindre une menace pour la vie ou la sécurité de tiers.

L'article 6-4 de la LCEN comporte une telle obligation de diligence en imposant aux opérateurs de plateformes en ligne de mettre en œuvre des procédures et des moyens humains et technologiques proportionnés permettant :

i) d'informer dans les meilleurs délais les autorités judiciaires ou administratives des actions mises en œuvre à la suite de la réception d'une injonction concernant un contenu à caractère haineux présent sur leur service ;
ii) d'accuser réception sans délai des demandes des autorités judiciaires ou administratives tendant à la communication des données dont ils disposent de nature à permettre l'identification des utilisateurs ayant mis en ligne des contenus illicites.

1. *Procédures et moyens humains et techniques permettant de traiter avec diligence les demandes des autorités publiques*

La capacité des opérateurs de plateformes en ligne à réceptionner les demandes des autorités publiques et à y répondre avec diligence contribue fortement à l'efficacité de la lutte contre la dissémination de contenus à caractère haineux en ligne.

La majorité des opérateurs, en l'espèce *Google*, *Meta*, *Snapchat*, *Yahoo*, *Microsoft* (*Bing*), *LinkedIn* et *Twitter*, déclarent avoir mis en place un canal de communication exclusivement consacré à la réception des demandes provenant des forces de l'ordre et des autorités judiciaires. Cette pratique paraît particulièrement utile pour garantir la traçabilité des demandes et l'authentification des demandeurs.

Google précise que les services enquêteurs qui utilisent l'interface qui leur est dédiée reçoivent systématiquement un accusé de réception dès l'envoi de la demande et peuvent suivre l'avancée du traitement des demandes en se connectant à leur dossier numérique.

Les opérateurs ayant mis en place un protocole de contact spécifique pour les forces de l'ordre précisent que les demandes qui leur parviennent par ce biais font l'objet d'une analyse par des équipes de modérateurs, généralement composées de juristes, visant à étudier la légalité des demandes préalablement à toute réponse.

Si la majorité des opérateurs affirment traiter avec diligence les demandes qui leur sont transmises par les interfaces de communication réservées aux forces de l'ordre, *Snapchat* se distingue en précisant traiter les demandes portant sur les infractions les plus graves telles que les atteintes imminentes à la vie ou à l'intégrité physique en 30 minutes.

Enfin, la *Fondation Wikimédia, Dailymotion* et *TikTok* déclarent avoir mis en place une adresse de contact permettant aux services enquêteurs de transmettre leurs demandes.

Préconisation :

- ❖ allouer des effectifs d'analystes proportionnés à l'exigence de traitement diligent des requêtes émanant des autorités publiques.

2. Signalement des suspicions d'infraction pénale aux autorités répressives

Aucun opérateur ne déclare avoir effectué un signalement de suspicion d'infraction pénale aux autorités françaises au cours de l'année 2022.

Toutefois, *Snapchat* précise être engagé de manière proactive avec les forces de sécurité des pays où il est actif et affirme signaler rapidement toute situation semblant présenter un risque particulièrement important pour des tiers (telles qu'une alerte à la bombe ou une menace d'attentat) aux agences de sécurité fédérales, lorsque le danger estimé se situe sur le territoire des États-Unis, ou à Interpol, lorsque la menace concerne une autre juridiction.

3. Réception et traitement des injonctions des autorités françaises

Les requêtes que les autorités françaises ont pu transmettre aux opérateurs dans le cadre de l'article 6-4 de la LCEN sont, d'une part, des injonctions portant sur des contenus à caractère haineux et, d'autre part, des demandes d'information visant à identifier les auteurs de contenus à caractère haineux.

Les réponses des opérateurs sont particulièrement hétérogènes à cet égard, certains étant en capacité d'identifier avec précision le nombre et la nature des demandes reçues concernant les contenus illicites en général, voire les contenus à caractère haineux plus précisément, ainsi que les actions prises à la suite de ces saisines, tandis que d'autres ne discriminent pas entre les signalements de contenus illicites transmis par des utilisateurs et les demandes d'information adressées par les services enquêteurs français sur l'année 2022.

i) Signalement de contenus illicites par les autorités françaises

Peu d'opérateurs affirment avoir été destinataires de signalements des autorités françaises, et, lorsque c'est le cas, les données sont très variables d'une plateforme à l'autre.

Ainsi *Twitter* indique avoir reçu un grand nombre de demandes de suppression de contenus de la part des autorités françaises (869). Comme indiqué plus haut, l'opérateur déclare un taux de décision de retraits de 43 % à la suite de ces demandes. Ce niveau relativement faible surprend, notamment au regard de son taux d'actions à la suite de signalements de signaleurs de confiance (66 %). L'Arcom invite *Twitter* à en examiner les raisons.

TikTok ne relève que 11 signalements pour l'année 2022. Là où *Meta* relève qu'aucun contenu ne lui a été signalé par les forces de l'ordre, *Bing* (destinataire de 155 demandes de retrait de contenus) signale à l'inverse que les signalements de contenus à caractère haineux qu'il a reçus (au nombre de 155) provenaient uniquement des « *autorités gouvernementales* » françaises, et non d'utilisateurs.

Parmi les opérateurs ayant reçu un nombre important de signalements portant sur des contenus à caractère haineux, *Google* (destinataire de 805 signalements de contenus sur *Google Search* et de 2 355 signalements de contenus sur *YouTube* uniquement via ses formulaires dédiés à l'article 6-4 de la LCEN), affirme être dans l'incapacité de distinguer les signalements de contenus provenant d'utilisateurs de ceux provenant des forces de l'ordre. L'Autorité note que les obligations de transparence prévues à l'article 15 du RSN imposeront notamment aux opérateurs de rendre public dans un rapport annuel le nombre d'injonctions reçues des autorités des États membre, classées par type de contenu illicite concerné, l'État membre qui a émis l'injonction ainsi que le délai médian nécessaire pour informer de sa réception l'autorité d'émission et pour donner suite à l'injonction.

Enfin, plusieurs opérateurs affirment n'avoir été destinataires d'aucune demande de retrait de contenus à caractère haineux de la part des autorités publiques ; c'est le cas de *Meta*, *LinkedIn*, *Snapchat*, *Yahoo* ou la *Fondation Wikimedia*.

ii) Demandes d'information

Les demandes d'information émises par les autorités judiciaires ou administratives françaises aux opérateurs recoupent en majorité les plateformes les plus ciblées par les demandes de retrait de contenus.

Les opérateurs sollicités ne déclarent pas de difficultés particulières pour répondre aux demandes des autorités judiciaires et administratives.

Les plateformes les plus ciblées par ces demandes, *Meta* et *Google*[25] en particulier, affirment répondre favorablement dans la majorité des cas, le taux de réponses positives dépassant 80 %, tandis que les opérateurs des plateformes *Yahoo*, *Bing* ou la *Fondation Wikimedia* affirment ne pas avoir été destinataires de ce type de demande.

À l'inverse, on relève que des opérateurs tel que *Snapchat* ou *TikTok* déclarent ne pas avoir été en mesure de quantifier le nombre de demandes d'information reçues.

[25] Google précise avoir reçu 6 017 demandes de communication de données de la part des autorités françaises entre janvier et juin 2022 et avoir répondu favorablement dans 85 % des cas.

III. Perspectives

A. Des opérateurs qui prennent progressivement la mesure de leur responsabilité sociale

Les grandes plateformes en ligne, et notamment les plus grands réseaux sociaux et moteurs de recherche, jouent un rôle essentiel dans l'accès à l'information, la participation au débat public et sa préservation ; nouvelles agoras, elles contribuent pleinement à la vitalité d'une société démocratique.

Pour autant, leur modèle économique repose largement sur la publicité et donc sur la capture, sans limite, de l'attention des usagers, et leur architecture de fonctionnement repose sur l'individualisation des contenus présentés (y compris publicitaires) en ayant recours à des traitements algorithmiques massifs des données individuelles collectées. Ces deux caractéristiques des très grandes plateformes et moteurs de recherche en ligne induisent des risques systémiques pour nos sociétés.

Ainsi, il existe un risque que ces plateformes ou moteurs de recherche accroissent artificiellement la visibilité des propos les plus conflictuels, facilitent la diffusion de contenus manifestement illégaux ou permettent la propagation virale de contenus préjudiciables, susceptibles d'accroître la conflictualité de l'espace public et de miner la cohésion de la société.

Par ailleurs, des problématiques connues de santé publique (par exemple, risque d'obésité, d'anorexie, de trouble du comportement, d'addiction aux services numériques) peuvent être amplifiées par les nouvelles dynamiques informationnelles découlant de l'usage des plus grandes plateformes ou moteurs de recherche.

Il convient donc d'identifier en amont, de constater en aval, d'atténuer et, plus généralement, de combattre ces risques systémiques induits pour nos sociétés lors de la conception même de ces services numériques pour préserver ces espaces nouveaux d'exercice de la liberté d'expression comme un bien commun.

Ces phénomènes sont désormais bien identifiés.

Ce bilan illustre le fait qu'un nombre croissant d'opérateurs, conscients de l'impact de leur service sur le fonctionnement de nos sociétés démocratiques, ont conçu, souvent avec un certain succès, des solutions pour modérer et atténuer les usages abusifs les plus manifestement préjudiciables de leurs services (dissémination de contenus à caractère pédopornographique ou terroriste, ventes de biens et produits illicites, incitation ou provocation aux atteintes aux biens) et anticiper, prévenir les effets induits sur nos dynamiques sociales, involontaires mais néanmoins réels, les plus préjudiciables, qui constituent autant de remise en cause de nos droits fondamentaux.

Ces mesures se sont toutefois développées dans un régime d'auto-régulation, sous la pression indirecte et informelle d'acteurs de la société civile ou des États, et s'avèrent désormais insuffisantes ou imparfaites pour répondre aux enjeux sociaux et démocratiques.

La mise en œuvre d'une démarche volontaire de responsabilisation renforcée, la conception de services intrinsèquement plus sécurisés pour l'usager et nos sociétés, l'anticipation des effets induits préjudiciables et le développement de mécanismes d'atténuation ont un coût qui peut être significatif et s'opposent en effet à la rationalité économique des plateformes dont le modèle dépend généralement de l'engagement maximal des utilisateurs. De la même manière, l'impératif de transparence, essentiel pour maintenir un niveau élevé de confiance dans nos espaces informationnels, trouve sa limite dans la volonté des plateformes de minimiser l'information partagée avec les publics sur le fonctionnement ou les effets de leur plateforme.

Ces limites appellent l'intervention d'un cadre de régulation, juridiquement opposable, mettant l'accent sur la transparence et la responsabilisation accrue des plateformes et moteurs de recherche en ligne. La France fait partie des quelques États membres de l'UE qui ont préfiguré ce cadre, induisant une responsabilisation accrue des plateformes en ligne vis-à-vis des risques de manipulation de l'information et de propagation de contenus haineux sur leur service. Chargée de veiller à la bonne application de ces lois nationales, l'Arcom a ainsi pu bâtir un dialogue exigeant avec les opérateurs et développer une expertise en matière de régulation des plateformes en ligne. Les bilans qu'elle a dressés depuis 2020 des moyens mis en place par les opérateurs rendent compte des premiers résultats encourageants obtenus.

C'est désormais un cadre collectif, le RSN, qui va s'appliquer à l'ensemble des intermédiaires numériques actifs en Europe, en accentuant les responsabilités pesant sur les très grandes plateformes et très grands moteurs de recherche en ligne à l'échelle de l'Union Européenne et mobilisant un réseau de régulateurs, travaillant de concert pour réguler ces acteurs systémiques par leur taille.

B. Le RSN consolide ces acquis communs et pose un cadre collectif de responsabilisation et de transparence

Le RSN renouvelle profondément le cadre juridique applicable aux acteurs de l'économie numérique.

Il élargit à l'ensemble des intermédiaires numériques des obligations qui étaient auparavant attendues des seules très grandes plateformes en termes de transparence, de diligence dans la modération des contenus illicites et de collaboration avec les tiers, tout en gardant une approche proportionnée, tenant compte des différences de nature et de fonction qui existent entre les acteurs au sein de l'écosystème numérique. Pour la catégorie nouvelle des plateformes en ligne, le RSN introduit un régime d'obligations où le niveau d'exigence varie en fonction de la taille des entreprises (les PME bénéficient ainsi d'un régime simplifié) et qui est particulièrement renforcé pour les acteurs qui induisent des risques systémiques par le nombre de citoyens européens qui y ont recours.

Ces nouvelles obligations contraignantes sont notamment :

i) **la transparence et la traçabilité** des injonctions de retraits de contenus ou de fourniture d'information émises par les autorités administratives et judiciaires en matière d'action contre les contenus illicites et d'identification des auteurs présumés ;

ii) **le renforcement des obligations en termes d'outils de signalement** des contenus illicites ou contraires aux conditions générales mis à la disposition du grand public ;

iii) **la création d'un statut de signaleurs de confiance** : cette pratique déjà existante de tiers de confiance en matière d'identification et de signalement de contenus illicites sera encadrée par un statut précis, reconnaissant leur expertise particulière et leur indépendance, et imposant aux opérateurs des obligations de diligence renforcées dans le traitement des notifications qui leur seront adressées par ces tiers. Ce statut de confiance s'accompagnera d'une transparence accrue de leur activité, notamment par la publication de rapports annuels ;

iv) **la protection des publics (notamment des mineurs)**, en leur donnant des outils pour participer, par leurs choix et leurs actions, à la sécurisation de l'espace numérique (information, paramétrage de la plateforme, signalement, recours) et de l'usage qu'ils en ont.

Afin de faire face à un opérateur qui ne respecterait pas ses obligations, le RSN prévoit un travail collectif de supervision des plateformes auquel tous les régulateurs participent dans une action concertée et, en cas de manquement avéré, pour le régulateur compétent (la Commission européenne ou l'autorité compétente du pays d'établissement selon les cas) un ensemble de moyens d'intervention allant d'un pouvoir d'enquête à la demande de restriction temporaire de l'accès au service auprès de l'autorité judiciaire, en passant par l'imposition des sanctions financières (ex. : amende et/ou astreinte journalière).

C. Pour les VLOPSEs, une meilleure prise en compte des risques systémiques

Le passage d'une forme d'auto-régulation éventuellement encadrée par de premières réglementations dans certains États membres à un cadre de régulation commun supervisé par la Commission européenne et mobilisant un réseau de régulateurs dans tous les États membres se justifie également par les risques particuliers que les très grands plateformes et moteurs de recherche en ligne sont susceptibles d'engendrer du fait de leur nombre d'usagers (supérieur à 45 millions sur un mois), leurs usages et de leur fonctionnement même.

Par leur ampleur et leur incidence, ces risques systémiques diffèrent de ceux imputables aux acteurs plus modestes. Les effets amplificateurs propres aux grands réseaux contribuent potentiellement à renforcer la diffusion de contenus illicites, ou leur capacité à induire ou amplifier des risques systémiques de nature à porter une atteinte durable et grave aux valeurs démocratiques, affaiblissant la qualité du discours civique, mettant en péril l'ordre public ou compromettant indument l'exercice de la liberté d'expression sur ces nouveaux espaces de débat public.

Cœur du projet de régulation européenne, l'identification, l'évaluation et l'atténuation de ces risques systémiques se traduisent par l'introduction d'instruments novateurs (approche par les risques et la mise en conformité, transparence généralisée, à l'initiative de la plateforme et de tiers indépendants de celle-ci, impliquant notamment des auditeurs, le monde académique et la société civile), déjà en germe dans certaines législations nationales mais qui se trouvent ici confortés et portés à l'échelle européenne.

Le RSN met ainsi l'accent sur les audits des risques par des auditeurs tiers indépendants, la mise en place de mesures appropriées pour atténuer ces risques et l'audit de ces mesures, l'accès aux données pour les experts de la lutte contre la dissémination de contenus préjudiciables (en particulier le monde académique) et, dans des circonstances exceptionnelles, la mise en place de mécanismes appropriés pour répondre aux crises, mesures d'urgence propres à faire face aux évènements extraordinaires entraînant une menace grave pour l'intégrité de l'Union ou d'une partie de celle-ci.

D. Eléments de calendrier et place de l'Arcom dans l'architecture européenne de régulation des plateformes en ligne

Ce cadre commun est amené à se développer et à s'enrichir dans le temps long qui s'impose en matière de libertés publiques.

Il sera notamment nourri par les échanges, d'une part, entre les régulateurs nationaux et le régulateur européen, dans le cadre d'une dynamique collective portée par un dialogue riche avec la Commission européenne et les États membres et, d'autre part, à l'échelle de chaque État membre, entre les régulateurs nationaux et l'ensemble des parties prenantes de cet État membre.

Ce travail a déjà débuté : la désignation des premiers VLOPSEs en avril 2023 par la Commission européenne marque une première étape. L'application des dispositions susmentionnées à ceux-ci à compter du 25 août 2023 est la seconde.

En France, les dispositions héritées de la loi du 24 août 2021 laisseront place à compter de début 2024 à l'entrée en vigueur des obligations du RSN s'appliquant à l'ensemble des intermédiaires numériques.

Cette entrée en vigueur est préparée par un travail conjoint des autorités compétentes en France telles que prévu par le projet de loi visant à sécuriser et réguler l'espace numérique (CNIL, DGCCRF, Arcom) et d'échanges soutenus avec les partenaires de l'Autorité au sein des pouvoirs publics (en matière de lutte contre les contenus haineux, la DILCRAH, PHAROS, le pôle national de lutte contre la haine en ligne, le secrétariat de la CNCDH et le CNNum), de la société civile, de la communauté académique et des plateformes.

Annexe 1

Liste des préconisations formulées par l'Arcom en matière de lutte contre la diffusion de contenus haineux en ligne

Sur la transparence en général

- **Préconisation n° 1 :** renforcer les moyens adéquats, notamment procéduraux, humains et technologiques, afin de satisfaire aux obligations de transparence avec diligence.

Sur l'accessibilité, la transparence et la clarté des conditions générales

- **Préconisation n° 2 :** rendre plus explicite, plus rapide et plus fluide le parcours permettant aux utilisateurs d'accéder aux conditions générales du service.
- **Préconisation n° 3 :** améliorer la lisibilité et les conditions de navigation au sein et entre les différentes pages consacrées aux conditions générales, aux règles communautaires, au centre d'aide (ou équivalents), etc.
- **Préconisation n° 4 :** préciser clairement, au sein des conditions générales du service, ce que recouvrent les contenus et comportements proscrits par le droit national et les règles de l'opérateur, et notamment l'interdiction des incitations à la haine et des comportements de harcèlement en ligne.
- **Préconisation n° 5 :** être le plus clair et explicite possible, au sein des conditions générales, sur l'existence et le fonctionnement du mécanisme interne de recours des décisions de modération tant pour les personnes qui signalent un contenu que pour celles dont le contenu est l'objet d'une modération.

Sur l'accessibilité et l'intelligibilité des dispositifs de signalement

- **Préconisation n° 6 :** dans les rapports de transparence, comptabiliser à la fois :
  - les contenus à caractère haineux signalés par le biais des formulaires *ad hoc* et via le dispositif de signalement à proximité directe des contenus ;
  - ceux retirés (i) sur la base des conditions générales et (ii) au titre du droit national.
- **Préconisation n° 7 :** améliorer l'accessibilité des dispositifs de signalement, notamment par des symboles et des intitulés plus explicites.
- **Préconisation n° 8 :** illustrer les intitulés des motifs de signalement par des exemples concrets.
- **Préconisation n° 9 :** rappeler, succinctement, les règles en vigueur sur le service en matière de contenus avant que l'utilisateur ne finalise son signalement.
- **Préconisation n° 10 :** offrir à l'utilisateur la possibilité de joindre à son signalement une description de tout élément qu'il jugera utile à l'analyse de celui-ci.
- **Préconisation n° 11 :** offrir à l'utilisateur la possibilité d'indiquer la ou les partie(s) du contenu que son signalement concerne.

- **Préconisation n° 12 :** permettre à l'utilisateur de signaler plusieurs contenus d'un même compte via un seul et unique signalement.
- **Préconisation n° 13 :** permettre à l'utilisateur signalant un contenu vidéo, d'indiquer un intervalle de temps complet et pas uniquement un horodatage de début.
- **Préconisation n° 14 :** offrir à l'utilisateur la possibilité d'indiquer si son signalement s'applique également aux liens inclus dans la description du contenu signalé.
- **Préconisation n° 15 :** être vigilant et, le cas échéant, maintenir une capacité de réaction face à des usages abusifs des outils de signalement.

En outre, conformément à ce qu'elle avait déjà indiqué dans le cadre de ses lignes directrices, l'Arcom rappelle la préconisation suivante :

- **Préconisation n° 16 :** permettre à l'utilisateur d'indiquer s'il souhaite être informé de l'évolution du traitement de son signalement.

Sur les moyens mis en œuvre pour la modération

- **Préconisation n° 17 :** renforcer la transparence des politiques de modération en rendant publics le nombre, la langue de travail et la localisation des modérateurs employés par l'opérateur.
- **Préconisation n° 18 :** assurer le dimensionnement adéquat des moyens humains dédiés à la modération des contenus illicites.
- **Préconisation n° 19 :** systématiser, dans les rapports de transparence, la distinction de l'origine du signalement entre utilisateurs, signaleurs de confiance et autorités publiques.
- **Préconisation n° 20 :** collecter et rendre publiques les données permettant d'objectiver la rapidité de traitement des notifications transmises par les signaleurs de confiance.
- **Préconisation n° 21 :** permettre à tout utilisateur de contester une décision de modération et assurer un égal traitement de ces recours, sans que soit exigée une argumentation juridique particulière.
- **Préconisation n° 22 :** lorsque le taux d'infirmation, à la suite d'un recours, des décisions ou actions de modération est élevé (comme c'est le cas pour *TikTok* et *Dailymotion*), prendre des mesures adéquates pour évaluer la pertinence de la modération initiales et y remédier le cas échéant.
- **Préconisation n° 23 :** allouer des effectifs d'analystes proportionnés à l'exigence de traitement diligent des requêtes émanant des autorités publiques.

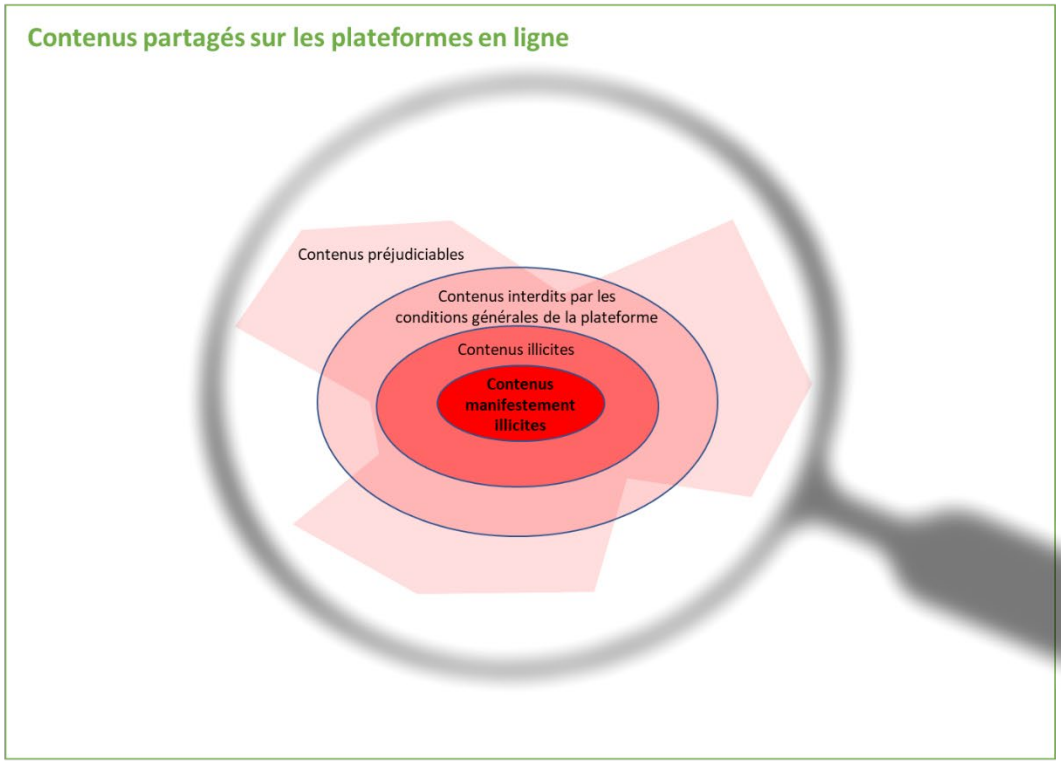
Annexe 2

La modération des contenus illicites et préjudiciables sur les plateformes en ligne

Les utilisateurs d'une plateforme en ligne sont tenus de partager uniquement des contenus qui ne sont interdits ni par la loi ni par les conditions générales (CG) de la plateforme, celles-ci étant souvent plus restrictives que la loi.

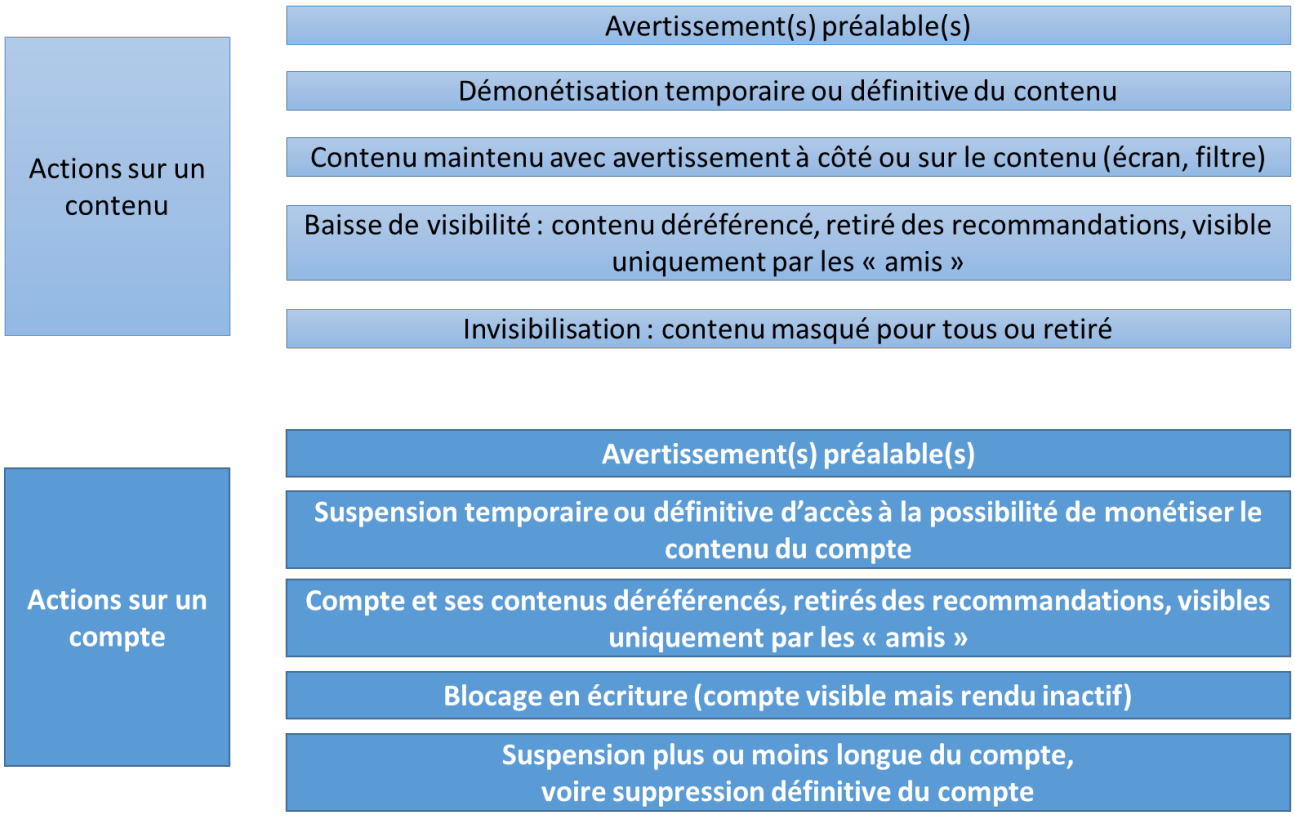
Lorsqu'une plateforme a connaissance d'un contenu ou d'un comportement susceptible d'être contraire à la loi ou à ses CG, à la suite d'un signalement ou parce qu'elle l'a elle-même détecté, elle doit l'examiner afin de déterminer s'il contrevient effectivement aux règles. Le contenu ou le comportement en question peut s'avérer manifestement illicite ou contraire aux CG.

Certains contenus ou comportements peuvent être agressifs, dérangeants ou désagréables pour tout ou partie des usagers qui les verront sans pour autant être proscrits par la loi ni les CG. Leur prolifération ou l'amplification de leur visibilité par la plateforme peut néanmoins contribuer à induire ou amplifier un risque systémique. Par exemple, la prolifération de contenus agressifs peut participer au sentiment de brutalisation des espaces numériques. Autres exemples : des comportements inauthentiques peuvent être susceptibles, à grande échelle, d'affaiblir la confiance dans discours civiques et les processus électoraux. La plateforme n'a pas nécessairement à retirer le contenu concerné ni à empêcher l'utilisateur de s'exprimer ; toutefois, certaines plateformes choisissent de limiter la viralité ou l'amplification algorithmique de ce type de contenus, en les retirant de leurs recommandations par exemple.



Les plateformes en ligne mettent en place différentes actions de modération qui peuvent porter, en fonction du cas et de la gravité, sur le contenu ou sur le compte. Le processus mis en œuvre dépend de chaque plateforme, libre de décider l'ordre et la gradation des interventions.

Dans un objectif de respect des droits des utilisateurs et pour un environnement numérique de confiance, le RSN impose toutefois aux plateformes d'exposer de façon claire et transparente ce processus à l'utilisateur et de lui permettre et lui faciliter l'exercice d'un recours en cas de désaccord avec une décision de modération.



Différents types d'actions de modération possibles *(non exhaustifs)*