OECD publishing

L'ÉTABLISSEMENT DE RAPPORTS DE TRANSPARENCE SUR LES CONTENUS TERRORISTES ET EXTRÉMISTES VIOLENTS EN LIGNE

UNE MISE À JOUR SUR LES 50 PRINCIPAUX SERVICES DE PARTAGE DE CONTENUS

OECD DIGITAL ECONOMY PAPERS

Juillet 2021 **No. 313**

OCDE

DES POLITIQUES MEILLEURES POUR UNE VIE MEILLEURE

Avant-propos

À la suite de sa 79e session tenue en juillet 2019, le Comité a établi l'ordre de priorité des travaux de suivi sur le thème des plateformes en ligne, compte tenu des options présentées dans le document publié sous la cote DSTI/CDEP(2019)7/REV1. L'un des deux projets prévus dans le cadre de l'actuel Programme de travail et Budget vise à élaborer un cadre et des indicateurs relatifs à l'établissement de rapports de transparence volontaires sur les contenus terroristes et extrémistes violents diffusés en ligne, l'autre consistant en une étude sur la portabilité des données (voir DSTI/CDEP/DGP(2019)2). Lors de la première phase du projet sur les contenus terroristes et extrémistes violents, deux rapports seront rédigés à un an d'intervalle afin de dresser un état des lieux des règles et procédures mises en place par les principales plateformes en ligne mondiales, ainsi que d'autres services de partage en ligne, face à ce type de contenus. Le premier rapport, intitulé Approches actuelles des 50 principaux services mondiaux de partage de contenus en ligne face aux contenus terroristes et extrémistes violents, a été publié en 2020. Le présent projet final du second rapport prend en considération les observations formulées par les délégués, à l'oral et par écrit, sur les premier et deuxième projets de rapport, ainsi que les retours des entreprises dont le profil a été établi à l'Annexe B.

Le projet sur les contenus terroristes et extrémistes violents est mené avec l'aimable soutien financier de l'Australie, de la Corée et de la Nouvelle-Zélande. Le Secrétariat tient à remercier Dr. Tomas Llanos pour son travail sur ce rapport.

Ce document ainsi que les données et cartes qu'il peut comprendre sont sans préjudice du statut de tout territoire, de la souveraineté s'exerçant sur ce dernier, du tracé des frontières et limites internationales, et du nom de tout territoire, ville ou région.

Note à l'intention des délégations :

Ce document est également disponible sur O.N.E, sous la référence :

DSTI/CDEP(2020)9/FINAL

Table des matières

Résumé

Les groupes terroristes et extrémistes violents utilisent Internet et les technologies connexes à des fins de radicalisation, de recrutement, de diffusion de propagande, de communication et de mobilisation. Les contenus terroristes et extrémistes violents publiés en ligne peuvent se répandre rapidement et à peu de frais, relayant des opinions dangereuses et touchant un vaste public. Le présent rapport est le second à examiner les règles et procédures mises en place par les 50 principaux services de partage de contenus en ligne face à ce type de contenus. Le premier rapport a fourni une base de comparaison, à partir de laquelle on a pu évaluer les évolutions pertinentes, exposées dans les présentes, notamment quant au nombre de services publiant des rapports de transparence sur les contenus terroristes et extrémistes violents.

Comme l'ont montré les attaques de Christchurch et de Halle, il est possible de diffuser en ligne des actes de violence terribles, choquants, en direct et sans filtre. Suite à de telles tragédies, des voix se sont élevées dans les forums internationaux (G20, 2019; G7, 2019; G20, 2017; Christchurch Call, 2019) pour intensifier les efforts visant à limiter la diffusion en ligne des contenus terroristes et extrémistes violents, d'une manière qui soit transparente, responsable et compatible avec les libertés et droits fondamentaux. Les acteurs du secteur des services en ligne ayant apporté leur soutien à l'Appel de Christchurch, par exemple, se sont engagés à « effectuer des rapports publics, réguliers et transparents, quantitatifs et reposant sur une méthodologie précise, sur la quantité et la nature de contenus terroristes et extrémistes violents détectés et retirés » (Christchurch Call, 2019). Dans leur Déclaration sur la prévention de l'utilisation d'Internet à des fins de terrorisme et d'extrémisme violent pouvant mener au terrorisme, les dirigeants du G20 réunis en 2019 à Osaka ont salué « l'engagement des responsables des plateformes en ligne à fournir régulièrement des rapports publics transparents » (G20, 2019). Cette transparence accrue améliorera la compréhension et l'évaluation des politiques et mesures adoptées par les services de partage de contenus face aux contenus terroristes et extrémistes violents, y compris en matière de modération des contenus. Elle contribuera également à garantir que les droits fondamentaux, tels que le droit à la vie privée, la liberté d'expression et le droit à un procès équitable, ne soient pas indûment restreints.

Ce second rapport d'analyse comparative examine dans quelle mesure les approches adoptées par les 50 plus grandes plateformes mondiales de partage de contenus en ligne face aux contenus terroristes et extrémistes violents ont évolué sur une période d'un an. À l'instar du premier rapport, cette nouvelle édition dresse un état des lieux objectif à un instant T. Il apporte également de la matière aux activités menées par l'OCDE en collaboration avec les pays membres, les entreprises, la société civile et les milieux universitaires pour concevoir un cadre multipartite, fondé sur le consensus, et un ensemble d'indicateurs pour les rapports de transparence volontaires que les entreprises publient sur les contenus terroristes et extrémistes violents diffusés en ligne. Le cadre et les indicateurs ont vocation à définir un modèle normalisé que toutes les entreprises souhaitant établir des rapports de ce type pourront utiliser, et que tous les membres de l'OCDE pourront promouvoir.

Les principales conclusions de ce second rapport d'analyse comparative sont les suivantes :

- D'une manière générale, le degré de transparence et de clarté des règles et procédures adoptées par les 50 principales plateformes face aux contenus terroristes et extrémistes violents s'est sensiblement accru. Depuis l'an dernier, six entreprises de services

supplémentaires ont commencé à établir des rapports de transparence consacrés spécifiquement à ce type de contenus.

- Les cinq services qui établissaient déjà des rapports sur les contenus terroristes et extrémistes violents fournissent à présent des informations supplémentaires, dont la nature varie toutefois.
- Les premiers rapports des nouveaux services à se prêter à l'exercice sont relativement élémentaires, si l'on excepte ceux de Microsoft. Il s'agit malgré tout d'un premier pas important dans la bonne direction, sans compter que leurs rapports pourraient s'étoffer au fil du temps.
- Malgré la hausse du nombre de services publiant des rapports de transparence, il existe encore un manque d'uniformité quant à leur définition des contenus terroristes et extrémistes violents et des concepts liés, à la teneur et la fréquence des rapports, ainsi qu'au mode de mesure ou de calcul des indicateurs qui y figurent. Par conséquent, il n'est pas encore possible à ce stade de dégager une perspective claire et complète, à l'échelle sectorielle, sur les effets des efforts déployés par ces services pour lutter contre la diffusion en ligne des contenus terroristes et extrémistes violents, notamment de leur impact sur les droits humains. La croissance continue du nombre de services établissant des rapports de transparence et la convergence accrue des indicateurs, des méthodologies et de la fréquence de ces rapports contribueront sans doute à faciliter l'émergence d'une telle perspective.
- De même, si le nombre de pays ou territoires ayant adopté des lois et réglementations liées aux contenus terroristes et extrémistes violents, ou qui envisagent de le faire, progresse, les textes restent, là encore, hétérogènes. Cette situation entraîne un risque de divergence des normes et exigences applicables à l'établissement des rapports.
- La pandémie de COVID-19 et les mesures de confinement qui l'ont accompagnée ont poussé certains services à davantage se reposer sur les systèmes de surveillance automatisés pour la détection et le retrait des contenus terroristes et extrémistes violents.
- Quatorze des 50 principaux services de partage de contenus en ligne étudiés dans ce rapport sont établis ou détenus par des sociétés mères établies en République populaire de Chine (ci-après dénommée la « Chine »), contre treize l'an dernier. L'un d'eux, TikTok, a rejoint les services qui établissent des rapports de transparence dédiés aux contenus terroristes et extrémistes violents, depuis la publication du premier rapport d'analyse comparative.
- Certains groupes terroristes et extrémistes violents se tournent vers des plateformes de moindre envergure qui ne disposent ni des ressources ni de l'expertise nécessaires pour gérer efficacement les contenus terroristes et extrémistes violents. Par conséquent, il serait utile de mener des travaux de recherche sur les services que ces groupes utilisent le plus, plutôt que de se limiter exclusivement aux 50 principaux services mondiaux. On pourrait également s'intéresser aux mesures que ces services prennent face aux contenus terroristes et extrémistes violents publiés sur leurs plateformes, ainsi qu'aux mécanismes de coopération et d'assistance par lesquels des plateformes davantage expérimentées pourraient aider les services plus modestes à combattre efficacement ce type de contenus.

Introduction

Si l'on ne peut contester les avantages qu'Internet a apportés dans nos vies, notamment sur le plan du développement du commerce transfrontalier, de la réduction des coûts de recherche et de transaction, et de l'émergence de nouveaux modes de communication, il a également fait naître de nouveaux défis. Et parmi eux, l'utilisation de services de partage de contenus en ligne pour publier et diffuser des contenus terroristes et extrémistes violents.

Les groupes terroristes et extrémistes violents ont fait la preuve de leur volonté de mettre à profit les nouvelles technologies à des fins de recrutement, de diffusion de propagande, de communication et de mobilisation (Office des Nations Unies contre la drogue et le crime, 2012). Outre qu'ils utilisent les réseaux sociaux pour donner de la résonance à leurs messages, ces groupes peuvent recourir à des applications de messagerie cryptée pour leur communication interne et la coordination d'attaques terroristes, tirant ainsi parti de la confidentialité qu'assurent ces applications pour éviter d'être détectés (Clifford & Powell, 2019). Comme l'ont montré les attaques de Christchurch et de Halle, l'amélioration de l'infrastructure des données mobiles a permis aux auteurs d'actes terroristes et extrémistes violents de diffuser leurs actions violentes sans filtre, sans censure et en temps réel, le public pouvant y accéder depuis un simple smartphone (Ahmed, 2020).

Face à ce phénomène, les géants des technologies tels que Google, Facebook, Twitter et Microsoft ont formé une alliance et pris un certain nombre de mesures pour mettre fin à la prolifération des contenus terroristes et extrémistes violents sur leurs plateformes en ligne, et empêcher l'utilisation abusive de leurs services. Comme indiqué dans la section 3 du premier rapport d'analyse comparative, plusieurs plateformes en ligne se sont unies pour créer le Forum mondial de l'Internet contre le terrorisme (Global Internet Forum to Counter Terrorism, GIFCT), un organisme désormais indépendant qui travaille de concert avec les pouvoirs publics, la société civile, les milieux universitaires et les organisations internationales pour lutter plus efficacement contre la prolifération des contenus terroristes et extrémistes violents en ligne. Ses efforts ont notamment débouché sur la création d'une base de données d'empreintes numériques – c'est-à-dire des empreintes composées d'une suite de lettres et de chiffres – de contenus terroristes et extrémistes violents connus, ayant déjà été retirés par au moins une des entreprises participantes. Cette base de données leur permet de choisir, selon leurs politiques respectives, d'empêcher rapidement toute nouvelle mise en ligne d'un même contenu ou de trouver des copies existantes de ce contenu sur les autres services participants. De même, un grand nombre de plateformes en ligne et d'autres services internet ont interdit explicitement l'utilisation de leurs technologies pour soutenir ou mener des activités terroristes et extrémistes violentes[1], et pris des mesures à la fois proactives et réactives pour empêcher et réduire au maximum les violations de leurs conditions générales (conditions d'utilisation) ou règles collectives (règles de la communauté). Ces mesures vont du simple avertissement à l'exclusion définitive de la plateforme concernée, en passant par le retrait des contenus visés ou la suspension du compte des utilisateurs[2].

Il existe toutefois de grandes disparités non seulement dans la façon dont les plateformes modèrent les contenus terroristes et extrémistes violents, mais aussi dans le degré de transparence dont elles font preuve à l'égard des contenus publiés sur leurs services et de leur traitement. Un autre aspect suscite des inquiétudes quant à la transparence et la responsabilité : la « modération de plus en plus agressive des contenus publiés par les utilisateurs » que les plateformes en ligne semblent pratiquer, qui pourrait porter atteinte aux droits et libertés fondamentaux des personnes, comme le droit au respect de la vie privée, à la

liberté d'expression et à un procès équitable (The Santa Clara Principles, n.d.). Pour y remédier, certaines entreprises technologiques se sont engagées à faire preuve de davantage de transparence dans la définition de leurs conditions d'utilisation et règles collectives, ainsi que dans leurs modalités d'application. Par ailleurs, plusieurs pays se sont engagés à travailler aux côtés des entreprises technologiques pour définir des stratégies visant à empêcher la mise en ligne et la diffusion de contenus terroristes et extrémistes violents sur les réseaux sociaux et les services de partage de contenus semblables. Ces engagements et promesses ont été exposés dans l'Appel de Christchurch (2019) et trouvent un écho dans un certain nombre d'appels internationaux à agir pour éradiquer de tels contenus, tout en veillant au respect des droits humains (G20, 2019) (G7, 2019).

À la suite des appels à l'action précités, l'OCDE a lancé un projet multidimensionnel en vue de concevoir un cadre de référence et de définir un ensemble d'indicateurs relatifs à l'établissement de rapports de transparence volontaires sur les contenus terroristes et extrémistes violents, sur la base d'un consensus international multipartite. Le projet prévoit notamment la publication, à un an d'intervalle, de deux rapports d'« analyse comparative ». Ils ont pour objet de dresser un état des lieux des règles et procédures mises en place par les 50 principales plateformes en ligne mondiales et autres services de partage de contenus en ligne (dénommés les « services ») afin d'identifier les points communs, l'évolution et les tendances des approches adoptées par ces services face à de tels contenus. Ils mettent l'accent sur la question de savoir si et, le cas échéant, dans quelle mesure les services établissent des rapports de transparence sur les contenus terroristes et extrémistes violents. Le premier rapport a été publié en août 2020 (OCDE). Le présent document (ci-après dénommé le « rapport ») correspond au second volet. Il examine l'évolution, au cours de l'année écoulée, des approches adoptées par les services pour lutter contre la diffusion des contenus terroristes et extrémistes violents. Comme le premier volet, il dresse un état des lieux objectif et factuel des règles et procédures actuelles des services en la matière. Il ne porte pas de jugement sur les mérites de ces règles et procédures, pas plus qu'il ne formule de recommandations. En revanche, il fournit une base factuelle permettant d'appréhender les approches adoptées par les services en question pour lutter contre les contenus terroristes et extrémistes violents et d'évaluer le degré de transparence et de responsabilité dont ils font preuve dans leur mise en œuvre.

Surtout, ce rapport apporte de la matière aux activités menées par l'OCDE, en collaboration avec les pays Membres, les entreprises, la société civile et les milieux universitaires, en vue de concevoir un cadre multipartite, fondé sur le consensus, et un ensemble d'indicateurs relatifs aux rapports de transparence volontaires que les services de partage de contenus consacrent aux contenus terroristes et extrémistes violents diffusés en ligne. Le cadre et les indicateurs ont vocation à être intégrés à un modèle normalisé que toutes les entreprises souhaitant établir des rapports de ce type pourront utiliser, et que tous les Membres de l'OCDE pourront appuyer.

La section 1 du présent rapport détaille la méthodologie de recherche mise en œuvre et son périmètre, et décrit les liens avec le premier rapport d'analyse comparative. La section 2 récapitule les conclusions phares du premier rapport et présente les principaux changements et développements observés, au cours de l'année écoulée, dans l'approche adoptée par les services pour lutter contre les contenus terroristes et extrémistes violents diffusés en ligne. La section 3 fait le point sur la structure et les initiatives du Forum mondial de l'Internet contre le terrorisme (Global Internet Forum to Counter Terrorism, GIFCT). La section 4 examine les principales évolutions, au cours de la dernière année, des propositions de législation ou de réglementation ayant trait aux contenus terroristes et extrémistes violents dans les pays de l'OCDE. L'Annexe A répertorie les 50 services mondiaux de partage de contenus en ligne les plus prisés. À l'Annexe B figurent les profils détaillés de ces services, l'accent étant mis sur leurs règles et procédures face aux contenus terroristes et extrémistes violents. Enfin, l'Annexe C propose un glossaire définissant les termes fréquemment rencontrés dans les rapports de transparence sur ces contenus.

1. Périmètre, méthodologie et plan de recherche

Le premier rapport d'analyse comparative (OCDE, 2020) consistait à recenser les règles, procédures et pratiques de lutte contre les contenus terroristes et extrémistes violents mises en œuvre par les 50 principaux services mondiaux. Ceux-ci comptent notamment des plateformes de réseaux sociaux, des services de communication en ligne, des plateformes de partage de fichiers, ainsi que d'autres services en ligne dont les activités permettent la mise en ligne, la publication, le partage ou le transfert de contenus numériques, ou facilitent l'échange de contenus vocaux, de vidéos et de messages ou d'autres types de communications en ligne. Comme indiqué à la section 1 du premier rapport, les services figurant dans cette liste des 50 premiers ont été sélectionnés sur la base de leurs chiffres de pénétration du marché ou de leur « popularité », en partant de l'hypothèse que les contenus terroristes et extrémistes violents diffusés sur des services populaires étaient davantage susceptibles de toucher un large public. Bien que le nombre de 50 soit nécessairement arbitraire, il a fallu déterminer une limite pour restreindre le champ des recherches effectuées pour ces rapports. Toutefois, il est important de noter que figurer sur la liste des 50 premiers services mondiaux n'équivaut pas nécessairement à figurer sur la liste des 50 principales plateformes aux plus hauts taux de présence de contenus terroristes et extrémistes violents. Ce rapport adopte la même approche, un an plus tard, afin d'identifier les évolutions observées au cours de l'année dans le traitement réservé par les 50 services les plus populaires aux contenus terroristes et extrémistes violents diffusés en ligne, pour les combattre. En particulier, il examine dans quelle mesure la définition des contenus terroristes et extrémistes violents par les services s'effectue avec plus ou moins de clarté, il décrit les procédures qu'ils suivent pour détecter ces contenus et les traiter, il indique si le nombre de services publiant des rapports de transparence sur les contenus terroristes et extrémistes violents a changé et précise quels indicateurs figurent dans ces rapports.

À l'instar du premier rapport, étant donné l'absence d'indicateurs communs susceptibles d'être utilisés pour déterminer la popularité des services étudiés, le rapport qui suit adopte une approche en deux phases pour définir quels services doivent entrer dans le périmètre des recherches. Dans un premier temps, les services ont été classés en trois catégories :

a. les réseaux sociaux, les services de vidéo en streaming et les services de communication en ligne ;
b. les services cloud de partage de fichiers ;
c. les « autres » services, catégorie qui comprend un service de gestion de contenus et une encyclopédie en ligne.

Au sein de chaque catégorie, les services les plus prisés ont été retenus à l'aide de la méthodologie suivante :

- Les plateformes de médias sociaux, les services de vidéo en streaming et les services de communications en ligne ont été identifiés en fonction du nombre d'utilisateurs actifs mensuels (UAM). Cet indicateur, couramment utilisé par les analystes du secteur et les investisseurs pour estimer la popularité et la croissance d'un service[3], permet de classer de manière relativement précise, selon leur taille relative, les services dont le succès repose sur l'engagement des utilisateurs.

- Les services cloud de partage de fichiers ont été identifiés d'après leurs parts de marché indicatives, une mesure couramment utilisée pour évaluer le poids d'une entreprise dans un secteur d'activité donné.
- La troisième catégorie comprend un service de gestion de contenus ainsi qu'une encyclopédie en ligne. La popularité de ces deux services ne peut pas être évaluée selon les mêmes critères que les deux autres, mais ils présentent toutefois un intérêt tel qu'il justifie leur inclusion. Leur poids a été évalué en fonction de données reflétant leur rayonnement et leur utilisation (part de marché indicative et nombre de pages vues mensuellement).

La liste des 50 principaux services mondiaux figure à l'Annexe A. Par rapport à la liste du premier rapport, elle ne présente que peu de changements. Si l'on excepte une fluctuation modérée de l'ordre de classement – TikTok et Telegram montant dans celui-ci, par exemple –, les seuls changements notables concernent l'inclusion de l'application chinoise de vidéos courtes Kuaishao en numéro 15 et la sortie du réseau social MySpace.

Les approches adoptées par les services dans la lutte contre les contenus terroristes et extrémistes violents diffusés en ligne ont été examinées en trois étapes. La première a consisté à appliquer le modèle de profil normalisé créé dans le cadre du premier rapport[4] pour dresser le profil de chaque service. Chaque service a fait l'objet d'un profil basé sur ses conditions d'utilisation, ses lignes de conduite et règles de la communauté, ses blogs, ses contrats de service ainsi que d'autres informations officielles (les « documents constitutifs »)[5], tous librement accessibles. Les services ont été contactés et disposaient d'un délai raisonnable pour formuler des observations sur l'exactitude de leur profil et fournir toute information complémentaire pertinente.

La deuxième étape a consisté à mettre à jour les profils à la lumière des réponses des services. Les versions finales des profils figurent à l'Annexe B.

La troisième a consisté à mettre à jour les principaux résultats du premier rapport avec les informations des profils des services issues de la nouvelle compilation. La section 2 du présent rapport fournit une vue d'ensemble factuelle et objective actualisée des approches adoptées par les 50 principaux services mondiaux dans la lutte contre les contenus terroristes et extrémistes violents diffusés en ligne.

La section 2 se concentre sur les changements et développements observés concernant les services :

a. les politiques concernant les notions de terroriste/terrorisme et extrémiste violent/extrémisme violent ;
b. la détection et le retrait des contenus terroristes et extrémistes violents, notamment les règles relatives au contrôle du respect des conditions générales d'utilisation du service, aux retraits et aux sanctions, et l'existence éventuelle de procédures de recours ;
c. les conséquences en cas de non-respect par les utilisateurs des conditions d'utilisation ou lignes de conduite et règles appliquées à la communauté en ligne ;
d. la publication volontaire de rapports de transparence sur les contenus terroristes et extrémistes violents, ainsi que leur contenu, la méthodologie utilisée et leur fréquence.

2. Approches des services en matière de lutte contre les contenus terroristes et extrémistes violents : points communs, progrès et tendances actualisés

Des différences persistent dans les descriptions des contenus terroristes et extrémistes violents et des concepts connexes, de même que dans les approches pour identifier les « organisations terroristes ».

Le premier rapport d'analyse comparative a permis d'identifier des approches divergentes dans les politiques adoptées par les services en matière de contenus terroristes et extrémistes violents et aux concepts connexes, ainsi qu'aux définitions qu'ils en donnent. Il en est de même concernant leur compréhension de ce que constitue un groupe ou une organisation terroriste, la fourniture d'explications détaillées et d'exemples constituant l'exception plutôt que la règle. Le tableau 1 montre qu'au cours de l'année passée, seuls des changements mineurs ont été observés.

Tableau 1. Approches adoptées par les services pour définir les contenus terroristes et extrémistes violents et les concepts connexes

Approche	1er rapport d'analyse comparative	2e rapport d'analyse comparative
Services définissant le terrorisme, l'extrémisme violent et les concepts connexes avec suffisamment de détails pour appréhender leur champ sémantique et fournissant des exemples lorsque cela est nécessaire	5[6]	6[7]
Services interdisant expressément l'usage de leurs technologies pour servir des objectifs terroristes ou extrémistes violents, utilisant (sans les expliquer en détail) les termes terroriste/terrorisme, extrémiste violent/extrémisme violent et des expressions similaires	19[8]	21[9]
Services plaçant les contenus terroristes et extrémistes violents dans la même catégorie que les discours haineux et/ou les contenus violents ou explicites	15[10]	13[11]

Services donnant des descriptions larges et/ou générales des comportements interdits, celles-ci pouvant être interprétées comme englobant les contenus terroristes et extrémistes violents.	16[12]	15[13]

Source : Annexe B dans (OCDE, 2020) ; Annexe B de ce rapport.

Certains services ont fait des efforts pour clarifier ce qu'ils considèrent comme des contenus terroristes et extrémistes violents ainsi que leur caractère inacceptable sur leur plateforme. En particulier :

a. Pinterest a actualisé ce qu'il considérait comme « des individus ou des groupes dangereux ».
b. Twitch a actualisé ses règles en matière de « Terrorisme et violence extrême », clarifiant sensiblement ce qu'il définit comme une organisation terroriste et comment ses équipes internes compétentes en matière de sécurité identifient ce genre de contenus. Ces clarifications ont élargi la définition des contenus qui entrent dans cette catégorie (y compris des formes de comportement auparavant considérées comme d'autres types d'abus).
c. Discord a édicté une nouvelle interdiction axée sur l'extrémisme violent, défini comme étant des « contenus dans lesquels des utilisateurs font l'apologie de la violence ou y apportent leur soutien en tant que moyen à visée idéologique ».
d. Microsoft a publié un *Digital Safety Content Report* (portant simultanément sur Skype et OneDrive), où il explique clairement que les « contenus tant terroristes qu'extrémistes violents sont interdits sur les plateformes et services Microsoft » et que le code de conduite contractuel des services Microsoft (*Microsoft Services Agreement Code of Conduct*) interdit la « publication de contenu terroriste ou extrémiste violent ».

Dans ses conclusions, le premier rapport d'analyse comparative indiquait que les services utilisaient des approches différentes pour identifier et définir une organisation terroriste. Un an plus tard, cette conclusion garde toute sa pertinence. Certains services, tels que Facebook et Instagram, utilisent leurs propres définitions des organisations terroristes en les distinguant des organisations promouvant la haine, des tueurs de masse ou en série, des groupes pratiquant la traite des êtres humains et des organisations criminelles[14]. D'autres services, tels que ceux fournis pas Microsoft, YouTube, Wordpress.com et Quora, se fondent sur la liste des organisations terroristes publiée par le gouvernement des États-Unis ou par les Nations Unies[15]. VK suit la définition légale de ce qui constitue un contenu terroriste dans les pays où il est présent[16]. La plupart des services ne fournissent toutefois aucune information à cet égard[17].

Les rapports de transparence traitant expressément des contenus terroristes et extrémistes violents sont encore rares parmi les 50 principaux services mondiaux, mais plusieurs explications sont possibles

Dans ses conclusions, le premier rapport d'analyse comparative faisait notamment remarquer que sur les 23 services ayant publié des rapports de transparence de quelque nature que ce soit, seulement cinq (Facebook, YouTube, Instagram, Twitter et Automattic) en avaient établi concernant spécifiquement les contenus terroristes et extrémistes violents. Au cours de l'année passée, Skype, OneDrive, Twitch, TikTok, Reddit et Discord ont rejoint le groupe des services fournissant des informations sur le retrait des contenus terroristes et extrémistes violents dans leurs rapports de transparence.

TikTok sort du lot dans la mesure où il a été le premier service chinois à publier des rapports de transparence et, à présent, à en publier spécifiquement sur les contenus terroristes et extrémistes violents. Cet effort s'est accompagné de la publication de nouvelles règles communautaires et du lancement d'un « centre de transparence » (*Transparency Center*) (Perez, TikTok to open a 'Transparency Center' where outside experts can examine its content moderation practices, 2020).

Il est important de noter que les acteurs criminels n'utilisent pas l'intégralité des services faisant partie des 50 principaux services mondiaux de partage de contenus en ligne pour diffuser des contenus terroristes et extrémistes violents, ce qui pourrait expliquer pourquoi les services ne publient pas tous des rapports de transparence consacrés spécifiquement à ces contenus. Par exemple, Pinterest, Medium et Meetup semblent être des lieux où les contenus terroristes et extrémistes violents n'abondent pas, voire n'apparaissent pas du tout, en tout cas à l'heure actuelle. De plus, les contenus terroristes et extrémistes violents ne sont pas répartis de manière égale ni même proportionnelle entre les services de partage de contenus en ligne où ils apparaissent. Ainsi, si certains services tels que 4chan, Telegram et YouTube détectent des volumes importants de contenus terroristes et extrémistes violents, leur nombre d'utilisateurs varie fortement. D'autres services, tels que Wikipédia et LinkedIn, jouissent certes d'une popularité élevée, mais semblent n'être utilisés que rarement pour publier des contenus terroristes et extrémistes violents.

Cependant, comme le montre la section 11 des profils répertoriés à l'Annexe B, des contenus terroristes et extrémistes violents sont apparus sur au moins 27 services à un moment ou à un autre, un chiffre à comparer aux seuls 11 services ayant publié à cette date des rapports de transparence qui y sont consacrés spécifiquement. Ici encore, des facteurs autres que l'absence de contenus terroristes et extrémistes violents peuvent expliquer pourquoi certains services n'établissent pas de rapports de transparence consacrés spécifiquement à ces contenus. Ainsi, 13 services sont des plateformes chinoises qui se voient empêchées de publier des rapports de transparence en raison des tensions existant entre les exigences légales locales et leur nature d'entreprises commerciales (voir plus loin, paragraphes 43-47). Il faut encore tenir compte de ce que les services proposant un chiffrement de bout en bout tels que iMessage/Facetime, Telegram et WhatsApp n'ont pas accès au contenu des communications de leurs utilisateurs ; raison pour laquelle ils ne peuvent établir de rapports de transparence consacrés spécifiquement aux contenus terroristes et extrémistes violents suffisamment détaillés[18].

L'absence ou la rareté relative de contenus terroristes et extrémistes violents, des contraintes réglementaires et des considérations techniques peuvent expliquer le manque de motivation ou l'impossibilité pour un service de délivrer des rapports de transparence sur ces contenus. Néanmoins, une plus grande clarté à cet égard ne pourrait qu'être bénéfique pour les efforts visant à établir des normes de traitement des contenus terroristes et extrémistes violents. Par exemple, si les entreprises déclaraient qu'elles ne publiaient pas de rapports de transparence sur les contenus terroristes et extrémistes violents parce qu'aucun contenu de ce type n'apparaît sur leurs services, il serait plus facile d'identifier les services jouant un rôle important dans la dissémination de tels contenus, le nombre de services entrant en ligne de compte étant ainsi réduit. Il est également possible que certains de ces services ne soient pas ceux comptant le plus grand nombre d'utilisateurs, c'est-à-dire ceux spécifiquement visés par ces deux rapports d'analyse comparative.

En effet, les spécialistes recommandent de ne pas limiter à quelques grandes plateformes les réponses à donner au problème des contenus terroristes et extrémistes violents en ligne. Au contraire, ils insistent pour que le problème soit considéré dans sa globalité, c'est-à-dire en étudiant de quelle façon la réaction des grandes plateformes induit des changements dans l'utilisation des services à des fins de dissémination de contenus terroristes et extrémistes violents, comme une migration massive vers des plateformes, services et applications moins en vue, y compris sur le *dark web* (Tech Against Terrorism, 2019). Les recherches ayant montré que certains groupes terroristes et extrémistes violents migrent vers de plus petites plateformes ne disposant pas des ressources et compétences nécessaires pour contrôler efficacement les

contenus terroristes et extrémistes violents (Tech Against Terrorism, 2019), il serait utile de faire porter les nouvelles recherches sur les services que ces groupes utilisent le plus pour diffuser des contenus terroristes et extrémistes violents. On pourrait également s'intéresser aux mesures que ces services prennent face aux contenus terroristes et extrémistes violents publiés sur leurs plateformes, ainsi qu'aux mécanismes de coopération et d'assistance par lesquels les plateformes en ligne bénéficiant d'une plus longue expérience pourraient aider les services plus modestes à combattre efficacement ce type de contenus.

Les rapports de transparence sur les contenus terroristes et extrémistes violents demeurent hétérogènes, mais sont plus complets

Le premier rapport d'analyse comparative a montré que les définitions utilisées et les types d'informations des cinq rapports de transparence sur les contenus terroristes et extrémistes violents publiés jusque-là différaient fortement. Un an plus tard, c'est toujours le cas. Cependant, une tendance générale commune aux cinq services ayant publié des rapports de transparence sur les contenus terroristes et extrémistes violents l'année passée se dessine : les informations fournies sont plus complètes.

Twitter, par exemple, en plus de signaler les comptes pour lesquels des mesures ont été prises ainsi que les comptes suspendus pour violation de ses règles, y compris celles visant à lutter contre le terrorisme et l'extrémisme violent, rapporte à présent des chiffres concernant les « contenus supprimés », c'est-à-dire le nombre d'éléments de contenu uniques (tels que des tweets ou des photos de profil, des bannières ou biographies) que Twitter a demandé à des titulaires de compte de retirer au titre qu'ils violaient les « règles de Twitter ». Twitter inclut également dans les données rapportées des tendances, dont certaines concernent les contenus terroristes et extrémistes violents. Ainsi, dans son dernier rapport, Twitter a observé une diminution de 9 % du nombre de comptes pour lesquels des mesures ont dû être prises au titre de la violation de sa politique en matière de terrorisme et d'extrémisme violent par rapport à la période précédente.

De même, en plus des indicateurs de ses rapports d'analyse comparative précédents, YouTube inclut à présent le nombre total de recours reçus par trimestre pour des vidéos retirées à la suite d'une violation des règles communautaires, ainsi que le nombre total de vidéos réactivées par YouTube, par trimestre, à la suite d'un recours après un retrait pour violation des règles communautaires. Il publie également le pourcentage de vidéos retirées avant qu'un utilisateur n'ait pu les voir et le pourcentage de retraits effectués après que les vidéos aient été vues (catégorie appelée « pourcentage de vues en infraction »).

Facebook rend compte des mêmes indicateurs que l'an dernier : 1) prévalence des violations liées à la propagande terroriste sur Facebook ; 2) nombre de contenus pour lesquels des mesures ont été prises par Facebook ; 3) pourcentage de contenus en infraction pour lesquels des mesures ont été prises par Facebook avant que des utilisateurs ne les signalent ; 4) nombre de recours contre des décisions prises à l'encontre de contenus spécifiques ; et 5) nombre de contenus que Facebook a rétablis après les avoir retirés. Instagram, de son côté, qui rendait compte des trois premiers indicateurs l'an dernier, les inclut à présent tous dans ses rapports. En outre, tant Facebook qu'Instagram font également état des tendances récentes concernant les contenus relevant de la haine organisée et du terrorisme pour lesquels des mesures ont été prises. Par exemple, le dernier rapport de transparence de Facebook note que les contenus relevant de la haine organisée pour lesquels des mesures ont été prises ont diminué de 4,7 millions d'éléments de contenu au premier trimestre de 2020 à 4 millions au deuxième trimestre de la même année, tandis que les contenus pour lesquels des mesures ont été prises au titre de la lutte contre le terrorisme ont augmenté de 6,3 millions au premier trimestre de 2020 à 8,7 millions au deuxième trimestre. Qui plus est, Facebook a mis à jour son document intitulé « Understanding the Community Standards Enforcement Report » (Comprendre le rapport sur la mise en application des règles communautaires) et fournit à présent des

explications extrêmement détaillées sur la façon dont tant Facebook qu'Instagram modèrent les contenus et calculent les indicateurs qu'ils mentionnent dans leurs rapports.

Les indicateurs rapportés par Automattic (maison mère de Wordpress.com) n'ont pas changé. Toutefois, dans le résumé proposé dans son rapport de transparence, Automattic rend à présent compte du nombre de sites ou contenus spécifiés dans les avis des unités de référence Internet (*Internet Referral Units, IRU)* pour la période du 1er janvier 2018 au 30 juin 2020.

Microsoft a commencé à publier un *Digital Safety Content Report*, lequel porte sur l'ensemble des produits et services de Microsoft destinés au public, entre autres OneDrive et Skype – des services dont le profil est établi dans les deux rapports d'analyse comparative – ainsi que d'autres produits tels que Bing, Xbox et Outlook. Microsoft inclut dans ce rapport des indicateurs concernant spécifiquement les contenus terroristes et extrémistes violents, dont le nombre de contenus pour lesquels des mesures ont été prises, le nombre de comptes suspendus en raison de contenus terroristes et extrémistes violents, le pourcentage de contenus traités que Microsoft a détectés, le pourcentage de contenus traités signalés par des utilisateurs ou des tiers, ainsi que le pourcentage de comptes suspendus ayant été rétablis à la suite d'un recours après un retrait au titre de la présence de contenus terroristes et extrémistes violents.

Twitch a publié son premier rapport de transparence en février 2021, celui-ci couvrant les premier et second semestres de 2020 (Twitch, 2020). Le rapport de transparence décrit les efforts consentis par Twitch ainsi que les méthodes mises en œuvre pour appliquer ses règles communautaires. Il fournit encore des informations sur la mesure dans laquelle des contenus terroristes et extrémistes violents apparaissent sur sa plateforme. Twitch explique qu'en raison de la nature de son service, à savoir la diffusion de flux vidéos en direct, la grande majorité des contenus sont éphémères. Le contenu en direct est signalé par un système de détection automatisé ou par des utilisateurs à l'équipe des modérateurs de contenu de Twitch (c'est-à-dire un personnel payé), laquelle prend des « sanctions » (consistant généralement en un avertissement ou une suspension temporaire de la chaîne) pour les cas de violations avérées. Si une violation s'accompagne de contenu enregistré, ce contenu est alors retiré. Il est à noter cependant que la plupart des sanctions n'entraînent pas le retrait du contenu, car, en dehors du signalement, il n'y a plus de traces de l'infraction. C'est pourquoi Twitch ne se repose pas sur la « suppression de contenu » comme principal moyen pour faire respecter ses règles communautaires. Le nombre de prises de « sanctions » est plus révélateur de ses efforts de mise en œuvre de ses règles communautaires. Ainsi, Twitch rapporte le nombre de sanctions prises pour des violations classées selon différentes catégories, dont l'une est intitulée « Terrorisme, propagande terroriste et recrutement ».

TikTok, Reddit et Discord ont également commencé à publier des rapports de transparence sur les contenus terroristes et extrémistes violents depuis la rédaction du rapport de l'année dernière. Ces premiers rapports sont toutefois modestes quant à leur portée :

a. TikTok rend compte du pourcentage de vidéos retirées pour violation de ses règles relatives aux propos haineux, au respect de l'intégrité et de l'authenticité et aux personnes et organisations dangereuses ;
b. Reddit fait état du nombre d'éléments de contenu relatifs à des organisations terroristes étrangères désignées (quoique la notion d'organisation terroriste étrangère désignée ne soit pas définie) qu'il a retirés au cours de la période couverte par le rapport ; et
c. Discord signale le nombre de suppressions de serveurs de contenu extrémiste violent par mois, lesquels sont détectés proactivement à l'aide d'outils automatisés.

Dans l'ensemble, les développements exposés ci-dessus, en particulier les informations fournies par Twitch dans son premier rapport de transparence, contribuent à mieux comprendre les efforts déployés par les services établissant des rapports pour lutter contre les contenus terroristes et extrémistes violents. De même, les règles mises à jour de Twitch sur le thème « Terrorisme et violence extrême » et le document actualisé de Facebook intitulé « Understanding the Community Standards Enforcement Report » montrent que ces acteurs prennent les contenus terroristes et extrémistes violents au sérieux, et reflètent leur engagement à tenir leurs utilisateurs informés sur la façon dont ils traitent le problème. En outre,

le fait que six nouveaux services fournissent des informations sur les contenus terroristes et extrémistes violents laisse entendre que le secteur reconnaît de plus en plus largement l'importance de la lutte contre ce phénomène ainsi que la nécessité de faire preuve de transparence en la matière.

Les chiffres rapportés par Twitch, Microsoft, TikTok, Reddit et Discord, en plus du nouveau décompte proposé par Twitter concernant les « contenus supprimés », mettent davantage en lumière la prévalence des contenus terroristes et extrémistes violents et, dans une certaine mesure seulement, témoignent d'une convergence à l'œuvre en matière de rapports de transparence, même si celle-ci demeure modérée. Toutefois, les chiffres rapportés par Microsoft le sont globalement pour tous les services et produits Microsoft destinés au grand public (sans que ceux-ci soient répertoriés en détail) et non individuellement par produit et par service. Cette approche permet d'obtenir des informations utiles au niveau de l'entreprise, mais sans perspective sur la répartition des contenus terroristes et extrémistes violents entre les produits et services de Microsoft. De plus, comme cela a été observé plus haut, les informations fournies par TikTok, Reddit et Discord sont relativement maigres et les approches des cinq autres services restent très hétérogènes. Il en résulte un potentiel de comparaison et d'analyse des différents rapports encore très limité. Toutes les observations notées dans cette section du premier rapport d'analyse comparative restent pertinentes et valides[19].

Compte tenu du faible nombre de services publiant des rapports de transparence et de l'hétérogénéité de leur contenu et des méthodes de calcul de leurs indicateurs, il n'est pas possible d'obtenir à l'heure actuelle une vision intersectorielle claire et exhaustive de la nature et de l'efficacité des mesures prises pour lutter contre les contenus terroristes et extrémistes violents diffusés en ligne, ou des conséquences de ces mesures sur le respect des droits humains. Accroître le nombre de services publiant des rapports de transparence sur les contenus terroristes et extrémistes violents et améliorer leur convergence tant en matière d'indicateurs que de méthodologies de calcul permettrait de mieux évaluer la nature générale et l'impact global des politiques des services et de leurs pratiques de modération pour lutter contre les contenus terroristes et extrémistes violents.

Modérateurs internes, utilisateurs modérateurs et outils automatisés : recours accru à l'automatisation à l'ère de la COVID-19

Le premier rapport a montré que les services adoptaient différentes approches pour modérer les contenus terroristes et extrémistes violents selon qu'ils mobilisaient le personnel interne, les utilisateurs ou des outils automatisés, ou une combinaison de ceux-ci. Le tableau 2 montre qu'il n'y a pas eu de changement significatif dans l'utilisation de ces approches au cours de l'année passée.

Tableau 2. Méthodes de modération des contenus par les services

Méthode	1er rapport d'analyse comparative	2e rapport d'analyse comparative
Services recourant à des modérateurs internes	40[20]	40[21]
Services recourant à des utilisateurs modérateurs	10[22]	10[23]

Services recourant à des outils automatisés	Au moins[24] 21[25]	Au moins[26] 23[27]

Note : Ces méthodes ne s'excluent pas mutuellement et peuvent donc être utilisées selon différentes combinaisons.

Source : Annexe B dans (OCDE, 2020) ; Annexe B de ce rapport.

Il convient de noter toutefois qu'à la suite de la pandémie de COVID-19 et des mesures de confinement, certains services, tels que Facebook, YouTube et Twitter, ont recouru de manière accrue à des systèmes de surveillance automatisés pour détecter et retirer les contenus problématiques, y compris les contenus terroristes et extrémistes violents. Ces systèmes ne sont pas capables des jugements nuancés parfois nécessaires pour déterminer si des contenus spécifiques correspondent bel et bien à des contenus terroristes et extrémistes violents (Duarte, Llanso, & Loup, 2017). Pour réduire la probabilité que de tels contenus passent à travers les mailles du filet, ces systèmes ont donc tendance à être programmés pour être plutôt trop prudents que pas assez. Par conséquent, cette tendance augmente le risque de faux positifs, comme reconnu par YouTube, qui a fait observer qu'en raison du recours accru à des systèmes de modération automatique, il était possible qu'un plus grand nombre de vidéos publiées par les utilisateurs et les créateurs soient retirées, y compris certaines vidéos qui ne contreviennent pas à [ses] règles. (YouTube, 2020). Certains commentateurs ont notamment observé que la part des contenus en langue arabe signalés comme terroristes était disproportionnée (Stokel-Walker, 2020). On le voit, le risque de confier la modération de contenus à des outils automatisés sans supervision humaine peut mener à exagérer le volume de contenus terroristes et extrémistes violents en ligne. D'un autre côté, on ne peut exclure que certains contenus terroristes et extrémistes violents échappent au repérage en raison de failles algorithmiques ou de l'efficacité des techniques de dissimulation mises en œuvre. Quoi qu'il en soit, l'ampleur des faux positifs et des faux négatifs reste incertaine, tout comme le fait que ces erreurs aient ou non une influence notable sur les indicateurs fournis par les services concernés.

Les limitations des outils automatisés dans la modération des contenus terroristes et extrémistes violents sont apparues au grand jour lors de l'attaque de Christchurch. Selon Facebook, la vidéo de l'attaque n'a pas déclenché ses systèmes de détection automatiques parce que le service ne disposait pas à ce moment de suffisamment de contenus représentant des séquences d'événements violents filmés à la première personne pour entraîner de manière efficace sa technologie d'apprentissage automatique. À la suite de ces événements, Facebook s'est mis à collaborer avec les autorités publiques et policières aux États-Unis et au Royaume-Uni pour obtenir des séquences filmées provenant de leurs propres programmes d'entraînement à l'utilisation d'armes à feu, constituant ainsi une source précieuse de données pour entraîner ses systèmes. Avec cette initiative, Facebook cherche à améliorer la détection de séquences vidéos d'événements violents réels filmés à la première personne tout en évitant de signaler par erreur d'autres types de vidéos, tels que les contenus fictionnels de films et de jeux vidéo (Facebook, 2019).

Persistance de l'hétérogénéité des mécanismes de notification, de sanction et de recours

La notification des décisions de sanction ainsi que la possibilité de former un recours sont importantes pour garantir un traitement équitable. Le premier rapport indiquait que les services présentaient des approches différentes en matière de notification et de recours. Le tableau 3 montre qu'un an plus tard, c'est toujours le cas.

Tableau 3. Approches des services en matière de notification et de recours

Approche	1er rapport d'analyse comparative	2e rapport d'analyse comparative
Les services ayant mis en place des mécanismes d'information des utilisateurs en cas de suspicion d'infraction à leurs conditions d'utilisation et autres documents constitutifs.	21[28]	23[29]
Services disposant de procédures de recours contre les décisions découlant de la modération des contenus et d'autres mesures prises en vertu de leurs documents constitutifs	23[30]	27[31]

Source : Annexe B dans (OCDE, 2020) ; Annexe B de ce rapport.

Les autres services soit ne disposent pas de procédures de recours, soit ne publient pas d'informations à cet égard.

Le premier rapport a aussi montré que, pour 22 services, il était difficile de déterminer avec exactitude s'ils procèdent à un examen a priori et/ou a posteriori des contenus pour vérifier le respect de leurs conditions d'utilisation et de leurs règles[32]. Après un an, ce nombre est tombé à 17[33].

Divulgation d'informations par les plateformes chinoises

Enfin, le premier rapport notait que les services chinois fournissaient généralement peu d'informations sur leurs pratiques et procédures en matière de modération de contenus au regard de leurs conditions d'utilisation et de leurs règles. Hormis TikTok, aucun d'entre eux n'a publié de rapport de transparence de quelque sorte que ce soit.

Comme évoqué plus haut, TikTok a commencé à établir des rapports de transparence sur les contenus terroristes et extrémistes violents. Il n'y a pas eu de changement dans les approches de publication d'informations des autres services chinois au cours de l'année passée[34].

Le premier rapport expliquait que le peu d'informations divulguées par les services chinois en matière de modération et de surveillance des contenus pouvait être dû au difficile équilibre à trouver entre, d'un côté, l'obligation pour ces services de se conformer aux réglementations locales (dans le cadre desquelles ils sont tenus de surveiller et censurer les contenus en étroite collaboration avec le gouvernement chinois) et, de l'autre, la nécessité de maintenir leurs services attractifs. Admettre qu'ils procèdent à une surveillance et à une modération des contenus pour se conformer aux réglementations locales pourrait porter atteinte à l'attractivité de leurs services au titre d'un défaut de confidentialité et de la censure. En dépit du peu d'informations divulguées, la surveillance et la censure assurées par les services Internet chinois en collaboration avec le gouvernement constituent un fait bien documenté. Ainsi, les études ont montré que des plateformes telles que WeChat mettent en œuvre différents outils de détection de mots clés et d'analyse pour contrôler les contenus et améliorer les mécanismes de surveillance chinois (Ruan, Knockel, Ng, & Crete-Nishihata, 2016). Les activistes politiques rapportent également avoir été suivis à la suite de propos tenus sur WeChat et que des enregistrements de dialogues instantanés ont déjà été produits comme preuve au tribunal (Zhong, 2018). Les plateformes et applications de réseaux sociaux jouent un rôle de tout

premier plan dans l'application du « système de crédit social » chinois, considéré largement dans les sociétés occidentales comme un système de surveillance et de contrôle gouvernemental de masse (Lix Xan Wong & Shields Dobson, 2019). Cette coopération est rendue obligatoire par un grand nombre de lois adoptées en vertu de la sécurité de l'État, de la sécurité publique, de la censure et de la taxation et qui ont donné au gouvernement chinois des pouvoirs étendus lui permettant d'accéder aux données du secteur privé générées en ligne par des entreprises exploitées en Chine (Wang, 2017).

Les préoccupations liées à la possibilité que ces services chinois fassent partie du système de surveillance du gouvernement chinois peuvent mettre un frein aux ambitions que pourraient avoir ces services d'étendre leurs activités à l'international. Le premier rapport notait que certains services chinois avaient consenti des efforts importants pour dissiper ce type de préoccupations à l'étranger. En particulier, TikTok a entrepris plusieurs initiatives pour augmenter sa transparence envers ses pratiques de modération et de retrait de contenus (Perez, TikTok to open a 'Transparency Center' where outside experts can examine its content moderation practices, 2020), en veillant à ce qu'elles n'entrent pas en conflit avec sa version chinoise, Douyin, et en assurant qu'aucune donnée des utilisateurs n'était envoyée en Chine (Cuthbertson, 2019). De la même façon, WeChat a adopté un modèle de censure suivant le principe « une application, deux systèmes » en vertu duquel seuls les utilisateurs WeChat dont les comptes sont couplés à des numéros de téléphone de Chine continentale sont surveillés et censurés (Ruan, Knockel, Ng, & Crete-Nishihata, 2016).

Toutefois, certaines sources semblent indiquer que les assurances de TikTok et WeChat sont trompeuses. Des études récentes ont montré que WeChat surveillait des comptes non enregistrés en Chine et utilisait les messages de ces comptes pour entraîner les algorithmes de censure à appliquer aux comptes enregistrés en Chine (Kenyon, 2020). De même, un livre blanc publié par la société de cybersécurité Penetrum a découvert que plus d'un tiers des adresses IP auxquelles le paquet APK de TikTok se connecte sont basées en Chine, concluant que « TikTok suit ses utilisateurs à un degré extrême et que les données collectées sont en partie, sinon entièrement, conservées sur des serveurs chinois du fournisseur d'accès Internet Alibaba » (Penetrum Security). Il a été signalé que les recherches effectuées sur TikTok révélaient un nombre de vidéos sur les manifestations à Hong Kong (Chine) bien moindre que ce qui était attendu, laissant entendre qu'une censure était à l'œuvre (Harwell & Romm, 2019). En outre, des lignes de conduite de modération promouvant la politique étrangère chinoise dans l'application TikTok ont fuité l'année passée (Hern, 2019). Sur la base de son utilisation de l'infrastructure chinoise et des liens étroits de sa société mère avec le parti communiste chinois, d'anciens ingénieurs spécialisés dans la sécurité de l'Organisation européenne pour la recherche nucléaire (CERN) ont averti récemment que TikTok était « l'outil de surveillance de masse et de collecte de données idéal pour le gouvernement chinois » (Kock, 2020). Ces préoccupations rendent encore plus souhaitable que des rapports de transparence fouillés soient publiés sur les pratiques de modération de contenus de ces services.

3. Le point sur le GIFCT

Le Forum mondial de l'Internet contre le terrorisme (GIFCT, *Global Internet Forum to Counter Terrorism*) a été fondé en 2017 par Facebook, Microsoft, Twitter et YouTube pour infléchir la diffusion de contenus terroristes et extrémistes violents sur les plateformes numériques. Pour une vue d'ensemble des objectifs et initiatives du GIFCT, veuillez vous reporter à la section 3 du premier rapport d'analyse comparative (OCDE, 2020).

Depuis sa fondation par Facebook, Microsoft, Twitter et YouTube, le Forum mondial de l'Internet contre le terrorisme s'est élargi à pas mesurés, mais sans discontinuer. Depuis l'adhésion d'Amazon, de Dropbox, de LinkedIn, de Pinterest, de WhatsApp et d'Instagram entre 2017 et 2019, Mega.nz, Mailchimp et Discord se sont joints également (GIFCT, 2021).

Dans sa nouvelle structure convenue en 2019, le GIFCT est administré par un comité opérationnel travaillant en étroite collaboration avec un large forum multipartite ainsi qu'un comité consultatif indépendant. C'est le comité opérationnel qui nomme son directeur exécutif, fournit le budget opérationnel initial et s'assure de la conformité de l'ensemble des activités du GIFCT à la mission qui lui a été confiée. Le comité opérationnel est composé comme suit :

- Membres fondateurs du GIFCT - Facebook, Microsoft, Twitter et YouTube
- Au moins une société du cadre élargi des membres, par roulement
- Nouvelles sociétés répondant à des critères de gouvernance et de direction
- Représentant du comité consultatif indépendant, par roulement, participant en qualité de membre sans droit de vote

La présidence tournante du comité opérationnel suit un cycle annuel. En 2020, la présidence du comité opérationnel est prise en charge par Microsoft (GIFCT, 2020).

Le forum multipartite comprend une grande diversité d'entreprises, de membres de la société civile et de gouvernements qui se sont engagés à faire respecter et à respecter les droits humains, et à empêcher l'exploitation des plateformes numériques par des terroristes. Servant de principal véhicule de partage d'information et d'échange d'idées, il a pour objectif d'aider à guider les activités du GIFCT ainsi que l'engagement de ses membres (GIFCT, n.d.).

Le 16 juin 2020, le GIFCT a annoncé que le premier comité consultatif indépendant était à présent membre à part entière du forum. Ses 21 membres comprennent des représentants de sept gouvernements, de deux organisations internationales et de 12 organisations de la société civile, présentant ensemble un domaine de compétences étendu.

Le Hash Sharing Consortium du GIFCT, qui met en commun des « empreintes numériques » d'images et de vidéos terroristes connues, gère une base de données comptant environ 300 000 empreintes uniques, soit quelque 250 000 images visuellement distinctes ainsi qu'approximativement 50 000 vidéos visuellement distinctes. Le consortium est composé de 13 sociétés jouissant d'un accès à la base de données partagée du secteur : Microsoft, Facebook, Twitter, YouTube, Ask.fm, Cloudinary, Instagram, JustPaste.it, LinkedIn, Verizon Media, Reddit, Snap et Yellow (GIFCT, 2020). Les empreintes sont étiquetées suivant la taxinomie suivante :

- Menace crédible imminente (ICT, *Imminent Credible Threat*) : publication d'une menace de violence spécifique, imminente et crédible à l'encontre de non-combattants et/ou d'infrastructures civiles.

- Violence explicite à l'encontre de personnes sans défense : meurtre, exécution, viol, torture ou atteintes graves à l'intégrité physique de personnes sans défense (exploitation de prisonniers, non-combattants évidents ciblés).
- Glorification d'actes terroristes (GTA, *Glorification of Terrorist Acts*) : contenus glorifiant, louant, justifiant ou célébrant des attaques ayant été perpétrées.
- Recrutement et formation (R&I, *Recruitment and Instruction*) : contenus visant à recruter des adeptes, les guider ou leur fournir une formation opérationnelle.
- Contenu lié à l'auteur de l'attaque en Nouvelle-Zélande : compte tenu de la viralité et de la diffusion entre les plateformes du manifeste de l'auteur et de la vidéo de l'attentat de Christchurch, et parce que les autorités néo-zélandaises ont déclaré illégal tout élément de ce type, le GIFCT a créé une banque de crise dans la base de données d'empreintes pour limiter la propagation de ces contenus.
- Contenu lié à l'auteur de l'attentat de Halle en Allemagne : c'est le 9 octobre 2019 que le GIFCT a activé pour la première fois depuis sa création son nouveau protocole de gestion des incidents liés aux contenus (CIP, *Content Incident Protocol*) suite à l'attaque terroriste de Christchurch en Nouvelle-Zélande, en mars de la même année. Le protocole de gestion des incidents liés aux contenus a été activé suite à la tuerie tragique de Halle, en Allemagne, et à la circulation de la vidéo de l'auteur de l'attaque sur de multiples plateformes numériques.
- Contenu lié à l'auteur de l'attaque de Glendale, Arizona, aux États-Unis : le 20 mai 2020, le GIFCT a activé son protocole de gestion des incidents liés aux contenus à la suite de la fusillade de Glendale, dans l'Arizona, ajoutant les empreintes de vidéos visuellement distinctes représentant le contenu de l'auteur pendant la fusillade (GIFCT, 2020).

En 2019, le Hash Sharing Consortium a mis en place une nouvelle mesure permettant à ses membres de contester les empreintes partagées dans la base de données. Si une société juge qu'une empreinte de la base de données a été ajoutée par erreur ou a reçu un libellé erroné, elle peut exprimer son désaccord de deux façons différentes. Premièrement, une société peut ajouter un libellé indiquant qu'elle reconnaît que l'empreinte correspond à un contenu terroriste, mais qu'elle pense que le libellé qui lui a été associé via la taxinomie est incorrect. Deuxièmement, une société peut ajouter à une empreinte un libellé indiquant que, selon elle, le contenu ne correspond pas à un contenu terroriste explicite (contenu contesté). Ces libellés sont visibles par toutes les sociétés faisant partie du Hash Sharing Consortium, de telle sorte qu'une société tierce soit en mesure de prendre sa propre décision quant à l'utilisation des empreintes associées à différentes catégories taxinomiques sur la base de ses propres processus et systèmes d'évaluation (GIFCT, 2020).

Le GIFCT a déclaré que son protocole de gestion des incidents liés aux contenus était un processus par lequel les sociétés membres du GIFCT détectent, évaluent rapidement et réagissent à un contenu potentiel circulant en ligne à la suite d'un événement terroriste ou extrémiste violent réel. Depuis la tragédie de Christchurch, les membres du GIFCT ont développé, affiné et éprouvé ce protocole. La procédure d'évaluation du protocole de gestion des incidents liés aux contenus a été lancée plus de 100 fois entre mars 2019 et novembre 2020 (GIFCT, n.d.).

Personne, pas plus un individu qu'une organisation, ne peut créer d'incident lié à un contenu. En effet, le protocole est basé sur l'existence d'un contenu diffusé en ligne à la suite d'un événement terroriste ou extrémiste violent réel – comme celui de Christchurch, de Halle ou de Glendale – et à la diffusion potentielle de ce contenu représentant l'action en direct d'un meurtre ou d'une tentative de meurtre produite par l'auteur d'une attaque ou un complice. Le protocole de gestion des incidents liés aux contenus constitue un processus à étapes multiples comprenant notamment la prise de décision de lancer la procédure correspondante, la communication de cette décision, l'évaluation des éléments constituant les contenus. Il comprend également d'autres étapes visant à informer les membres du GIFCT et les gouvernements concernés des contenus liés à des événements réels qui pourraient se

manifester en ligne. Un protocole de gestion des incidents liés aux contenus prend fin lors de sa « conclusion » officielle à l'initiative des membres fondateurs du GIFCT dès lors que le volume de contenus a baissé sensiblement sur l'ensemble des plateformes des membres du GIFCT (GIFCT, n.d.).

Le GIFCT a également créé plusieurs groupes de travail, dont l'un est consacré à la transparence. Le groupe de travail sur la transparence s'est réuni mensuellement depuis juillet 2020 et est comprend des représentants des entreprises, des gouvernements, d'organisations internationales, du monde académique et de la société civile. Son programme de travail vise à améliorer la compréhension et l'utilité de la transparence.

Enfin, au cours de la dernière année, le GIFCT a adapté son programme de partage d'URL dans le cadre d'un projet pilote de partage d'adresses URL de 12 mois mené avec SITE Intelligence, une société proposant des abonnements à des services de surveillance et d'analyse de contenus terroristes et d'autres types d'atteintes en ligne. Le projet pilote a permis à plusieurs membres plus récents du GIFCT d'accéder au flux source de SITE, lequel donne accès à un tableau de bord fournissant plus de contexte sur une URL donnée, y compris les liens organisationnels associés aux contenus terroristes et la traduction du contenu en anglais, outre d'autres aides contextuelles. Dans le cadre de ce programme, le GIFCT a partagé aujourd'hui près de 24 000 adresses URL depuis son lancement (GIFCT, 2020).

4. Législations et réglementations liées aux contenus terroristes et extrémistes violents en vigueur ou à l'étude

Les réseaux sociaux et autres services de communication en ligne ont été identifiés[35] comme constituant une panoplie complète d'outils pour le recrutement, l'engagement et la coordination des groupes terroristes et extrémistes violents. De plus, les informations partagées sur ces plateformes sont perçues, par les personnes susceptibles de tomber dans l'extrémisme, comme plus fiables que celles des médias d'actualités classiques, car elles ne seraient pas cadrées par les biais perçus des médias[36].

En raison de l'usage abusif que les groupes terroristes et extrémistes violents font des services en ligne dans le but de disséminer des informations de propagande et de recrutement, les entreprises technologiques ont fait l'objet d'une pression accrue de la part des gouvernements et des institutions du monde entier pour intensifier leurs efforts de lutte contre les activités de ces groupes. Devant l'insuffisance des efforts consentis jusque-là par le secteur pour lutter contre les contenus terroristes et extrémistes violents, certains gouvernements ont commencé à proposer et à adopter des législations et réglementations, ou à mettre en œuvre d'autres stratégies pour endiguer la propagation en ligne de ces contenus. Du reste, les réponses législatives et réglementaires aux contenus terroristes et extrémistes violents varient tout autant que les politiques et approches adoptées par les services en ce qui concerne les rapports de transparence sur ces contenus. Il y a donc un manque de coordination des deux côtés. Cette section offre une vue d'ensemble des législations et réglementations liées aux contenus terroristes et extrémistes violents qui ont été adoptées ou qui sont à l'étude aujourd'hui.

Australie

À la suite des attaques terroristes de Christchurch, le parlement australien a réagi en adoptant une loi, la *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019* (la « loi »), entrée en vigueur le 6 avril 2019 (Gouvernement de l'Australie, 2019). La loi introduit dans le code pénal de nouvelles infractions relatives aux contenus violents odieux diffusés en ligne.

Ceux-ci portent sur tous contenus audio, visuels ou audiovisuels enregistrant ou diffusant un comportement violent et odieux produits par le ou les auteurs de ce comportement (ou un complice) et qu'une personne raisonnable considérerait comme choquants au vu des circonstances. Les comportements violents odieux sont définis comme étant le meurtre ou la tentative de meurtre, l'acte terroriste, la torture, le viol ou l'enlèvement. Il n'est pas nécessaire que la personne soit reconnue coupable d'une infraction pour que son comportement constitue un comportement violent odieux. Aux fins de la loi, la question de savoir si le contenu violent odieux a été modifié ou non (par exemple par superposition d'un autre contenu) n'a pas d'importance. Toutefois, si le matériel est modifié (par un travail d'édition approprié) au point qu'il ne répond plus aux critères du matériel violent odieux, il sort du champ d'application de la législation.

En vertu de la loi, un fournisseur de services Internet, de services de contenu ou de services d'hébergement qui ne signalerait pas à la police fédérale australienne (AFP, *Australian Federal Police*) « dans un délai raisonnable » du matériel violent odieux qu'il sait être accessible à travers ou sur son service commet une infraction si le comportement sous-jacent s'est produit ou se produit en Australie. L'expression « délai raisonnable » n'est pas définie par la loi.

Toutefois, la note d'explication indique que la question est laissée à l'appréciation ultime de l'instance de jugement (par exemple un jury) et qu'elle dépendra de différents facteurs, tels que le volume du matériel (par exemple la fréquence selon laquelle il a été publié et republié) ainsi que les capacités et les ressources du fournisseur de services (c'est-à-dire la capacité technique dont il dispose pour retirer le matériel concerné).

En outre, la loi dispose qu'un fournisseur de services de contenu ou d'hébergement qui ne retirerait pas diligemment du service de contenu ou d'hébergement du matériel violent odieux susceptible, selon toute vraisemblance, d'être accessible en Australie (indépendamment de l'endroit où se trouve le service proprement dit) commet une infraction. La question de savoir si oui ou non le contenu spécifique a été « retiré diligemment » est ici encore laissée à l'appréciation de l'instance de jugement des faits et dépendra de facteurs tels que le type et le volume du matériel et les capacités et ressources du fournisseur de services.

La loi habilite également le commissaire à la sécurité électronique (*eSafety Commissioner*) à émettre des avis à l'intention des fournisseurs de services de contenu ou d'hébergement pour les avertir de ce que leurs services pourraient être utilisés, au moment de l'émission de l'avis, pour accéder à des contenus violents odieux. Un tel avis ne crée toutefois pas de responsabilité sur le plan pénal. Dans une procédure qui s'ensuivrait, l'avis constituerait un élément de présomption indiquant que le fournisseur de services a fait preuve d'imprudence en laissant son service être utilisé pour accéder à du matériel violent odieux. Des poursuites resteront nécessaires pour prouver les éléments constitutifs de l'infraction conformément aux règles régissant l'administration de la preuve en droit pénal. La présomption peut être réfutée par le fournisseur s'il fournit des éléments de preuve du contraire.

Le commissaire à la sécurité électronique peut faire usage d'un pouvoir d'instruction en vertu du paragraphe 581(2A) de la loi australienne de 1997 sur les télécommunications (*Telecommunications Act 1997*) afin de transmettre aux fournisseurs de services des instructions écrites en rapport avec un pouvoir ou une fonction quelconque de celui-ci. En juillet 2019, le ministre des Communications, de la cybersécurité et des arts a conféré au commissaire à la sécurité électronique, par la voie d'un instrument législatif, une nouvelle fonction en vue de promouvoir la sécurité en ligne des Australiens en les protégeant d'un accès ou d'une exposition à du matériel qui promeut des actes terroristes ou des crimes violents, qui incite à en commettre ou qui en ordonne l'exécution.

Le commissaire à la sécurité électronique a fait usage du pouvoir d'instruction en septembre 2019 afin de formaliser une mesure déjà prise par les fournisseurs de services Internet en vue de bloquer les sites connus donnant accès à des séquences filmées des attaques de Christchurch et/ou au manifeste de l'auteur. L'instruction de blocage était une mesure temporaire prise à la suite d'une procédure de consultation du secteur qui donnait l'occasion aux administrateurs de sites internet de retirer volontairement le contenu visé. L'instruction a expiré en mars 2020.

Par ailleurs, le groupe de travail australien de lutte contre les contenus terroristes et extrémistes violents diffusés en ligne (le « groupe de travail ») a été mis sur pied en mars 2019 avec pour objectif de fournir au gouvernement des conseils sur des mesures pratiques, concrètes et efficaces, ainsi que sur les obligations en matière de lutte contre la mise en ligne et la diffusion de matériel terroriste et extrémiste violent (Department of the Prime Minister and Cabinet, 2019). Conformément à ses attributions, le groupe de travail a publié le 30 juin 2019 un rapport identifiant des actions et des recommandations relevant de cinq champs d'action identifiés : la prévention, la détection et le retrait, la transparence, la dissuasion et le renforcement des capacités. Ces actions et recommandations comprennent notamment les points suivants :

a. Les plateformes numériques doivent continuer à mettre au point des solutions techniques visant à prévenir la mise en ligne sur leurs services de matériel terroriste et de extrémiste violent, et informer le gouvernement australien de l'avancée de ces solutions.
    - Des représentants du secteur ont rendu compte des développements au gouvernement australien en septembre 2019 et lui ont remis des rapports de mise

en œuvre esquissant les mesures qu'ils comptaient prendre pour appliquer les recommandations du groupe de travail. L'échéance suivante pour la soumission des rapports annuels des représentants du secteur au gouvernement australien était fixée en novembre 2020.

b. Les plateformes numériques doivent collaborer avec les autres membres du GIFCT pour enrichir la base de données de partage d'empreintes et le consortium de partage d'adresses URL pour faire converger autant que possible les catégories de contenus violents interdits par les plateformes conformément à leurs normes communautaires et conditions d'utilisation respectives, telles que celles relatives aux contenus violents et explicites ou aux contenus à caractère sanglant.

c. Les plateformes numériques doivent disposer de mécanismes de recours clairs et efficaces permettant aux utilisateurs de contester les décisions de modération en matière de matériel terroriste et extrémiste violent.

d. Sous la supervision et la direction du comité de contre-terrorisme d'Australie et de Nouvelle-Zélande (l'*Australia-New Zealand Counter-Terrorism Committee*), les plateformes numériques et les agences publiques australiennes concernées devaient organiser un « événement test » en 2019-2020 simulant un scénario dans le cadre duquel toutes les parties auront l'occasion de juger si les outils industriels et les procédures gouvernementales fonctionnent correctement, en particulier dans leur évolution en réaction aux technologies et aux investissements accrus dans la modération de contenu.

   - Cet événement d'épreuve a eu lieu le 1er octobre 2020 et a permis de parachever les dispositions de gestion des incidents liés aux contenus en ligne de l'Australie (*Online Content Incident Arrangement*).

e. En concertation avec le groupement Communications Alliance Ltd, le commissaire à la sécurité électronique a pour charge d'élaborer un protocole visant à régir l'utilisation temporaire du pouvoir du commissaire d'ordonner aux fournisseurs de services Internet de bloquer les sites hébergeant des contenus incriminés en cas d'événement de crise en ligne.

   - Ce protocole a été finalisé en décembre 2019.

f. Le gouvernement australien devrait poursuivre son travail législatif pour fixer un cadre légal au blocage de contenus à caractère terroriste et extrémiste violents en ligne durant les événements de crise.

   - En décembre 2019, le gouvernement australien a annoncé une nouvelle proposition de loi sur la sécurité en ligne (*Online Safety Act*) qui inclurait une nouvelle mesure de blocage de contenus. Dans le cadre de cette loi, le commissaire à la sécurité électronique se verrait accorder le pouvoir d'ordonner aux fournisseurs de services Internet de bloquer des domaines contenant du matériel terroriste ou extrémiste violent pendant une période limitée dans le temps au cours d'un événement de crise en ligne. Ce pouvoir serait plus ciblé que le pouvoir de blocage existant du commissaire à la sécurité électronique et conférerait aux fournisseurs de services Internet une immunité sur le plan civil dans la mesure où ils agiraient conformément à une instruction de blocage. Le gouvernement australien intègre dans son travail les commentaires issues de la consultation publique menée à ce sujet à mesure qu'il avance dans l'élaboration de son projet de réforme de la législation relative à la sécurité en ligne[37].

g. Les plateformes numériques doivent publier (au moins une fois par semestre) des rapports exposant les efforts mis en œuvre pour détecter et supprimer les contenus terroristes et extrémistes violents sur leurs services. Destinés à montrer la nature et la portée des mesures prises par les plateformes, ces rapports pourraient inclure les informations suivantes :

- le nombre d'éléments signalés par les utilisateurs comme susceptibles d'enfreindre les règles contre la promotion des contenus terroristes ou extrémistes violents ;
- le nombre total d'éléments retirés par la plateforme ;
- le nombre et le type (vidéo, chaîne, par exemple) des éléments de contenu terroriste et extrémiste violent retirés par les plateformes ;
- des exemples de contenus signalés au titre de la promotion du terrorisme ou de l'extrémisme violent qui enfreignaient et n'enfreignaient pas les lignes de conduite de la plateforme ;
- le nombre d'éléments de contenu terroristes et extrémistes violents signalés ou identifiés par les systèmes de la plateforme ;
- le nombre total d'éléments de contenu terroriste et extrémiste violent ayant fait l'objet d'une modération, répartis en fonction de l'entité à l'origine du signalement : utilisateurs, systèmes, autres sources, et le volume total des contenus retirés ;
- le temps moyen nécessaire à l'examen des éléments signalés et à la prise de mesures à leur encontre ou le nombre de visionnages de contenus terroristes et extrémistes violents par les utilisateurs avant qu'une mesure soit prise ;
- la mise en œuvre de contrôles appropriés sur les flux en direct afin de réduire le risque que des utilisateurs disséminent du matériel terroriste et extrémiste violent en ligne (Department of the Prime Minister and Cabinet, 2019).[38]

Canada

L'approche actuelle du Canada en matière de poursuites judiciaires à la suite de la publication de contenus terroristes et extrémistes violents est fondée sur le Code criminel canadien (*Criminal Code*). Le Canada a défini plusieurs infractions pénales en lien avec les comportements en ligne dangereux. Il prohibe notamment la propagande haineuse consistant à défendre ou à promouvoir les génocides, l'incitation à la haine dans l'espace public susceptible de porter atteinte à la paix, ainsi que l'incitation intentionnelle à commettre des infractions relevant de la haine et du terrorisme. Le Code criminel confère également aux cours et tribunaux le pouvoir d'ordonner le retrait de contenus déterminés des services Internet hébergés au Canada. Ces procédures de retrait existent en lien aux interdictions ci-dessus.

Le Canada est en train de réviser son approche des contenus terroristes et extrémistes violents sur les plateformes de réseaux sociaux. En 2019, le ministre du Patrimoine canadien a été chargé de « créer de nouveaux règlements pour les plateformes de réseaux sociaux, en commençant par exiger que toutes les plateformes éliminent le contenu illégal, y compris le discours haineux, dans les 24 heures, sous peine de sanctions importantes. Cela devrait inclure d'autres préjudices en ligne tels que la radicalisation, l'incitation à la violence, l'exploitation des enfants ou la création ou la diffusion de propagande terroriste. » (Gouvernement du Canada, 2021)

Dans ce contexte, il est rapporté que le ministre du Patrimoine canadien Steven Guilbeault travaille actuellement à l'introduction d'une législation en vue de créer un nouvel organe officiel de régulation habilité à surveiller les plateformes de réseaux sociaux et à imposer des amendes aux entreprises de réseaux sociaux qui permettraient à des contenus, comme des discours de haine, de rester sur leur plateforme (Thompson, 2021). Le gouvernement a également l'intention d'introduire une mesure prévoyant l'émission d'un avis de retrait dans les 24 heures, qui donnerait au régulateur le pouvoir de contraindre les plateformes à retirer les contenus que le régulateur juge illégaux ou haineux, ou qui encouragent autrement la radicalisation, incitent à la violence ou promeuvent la propagande terroriste (Patriquin, 2021).

Ces nouvelles législations sont en cours d'élaboration, l'objectif étant d'introduire de nouvelles lois et règles applicables aux plateformes de réseaux sociaux en 2021. L'approche du Canada tient compte des évolutions actuelles et des régimes réglementaires existants dans le monde, afin de garantir la sécurité et le bien-être en ligne des Canadiens tout en préservant la liberté d'expression.

Union européenne

Le 30 septembre 2020, la Commission européenne a adopté un rapport évaluant les mesures prises par les États membres pour se conformer à la directive de l'UE sur la lutte contre le terrorisme (Commission européenne, 2020), y compris en ce qui concerne son article 21, qui oblige les États membres à prendre les mesures nécessaires pour faire rapidement supprimer ou bloquer les contenus en ligne hébergés sur leur territoire et constituant une provocation publique à commettre une infraction terroriste. Ces mesures doivent être transparentes et fournir des garanties suffisantes (y compris en matière de réparation judiciaire) pour veiller à ce qu'elles soient limitées à ce qui est nécessaire et proportionné, et que les utilisateurs soient informés de la raison de ces mesures. Dans l'ensemble, la transposition de cet article ne s'est pas faite de manière homogène entre les États membres, certains choisissant de ne pas transposer intégralement cet article 21 dans leur législation nationale.

À la suite d'une proposition de la Commission européenne en septembre 2018, le Parlement européen et le Conseil de l'Union européenne se sont mis d'accord sur le « règlement relatif à la prévention de la diffusion de contenus à caractère terroriste en ligne » en décembre 2020. Le règlement a été adopté durant la session plénière du Parlement européen le 28 avril 2021. Les obligations qui y sont définies s'appliquent aux fournisseurs de services d'hébergement établis dans l'Union européenne afin de prévenir l'utilisation abusive de leurs plateformes par des terroristes. Les autorités nationales compétentes seront habilitées à transmettre des instructions directement aux entreprises pour leur demander de retirer des contenus dans l'heure de la réception d'un ordre de retrait. Les États membres peuvent également exiger que les entreprises prennent des mesures proactives lorsque les mesures existantes ne sont pas suffisantes pour diminuer efficacement les risques que des contenus à caractère terroriste soient diffusés sur leurs services. Les fournisseurs de services d'hébergement seront libres de choisir les mesures qu'ils considèrent les plus appropriées compte tenu de leur taille, de leurs capacités et des ressources disponibles.

La définition de ce que constituent des contenus à caractère terroriste en ligne recoupe celle des infractions terroristes qui est exposée dans la directive relative à la lutte contre le terrorisme, celle-ci englobant les contenus les plus dangereux, y compris les contenus incitant à commettre des actes terroristes ou défendant ceux-ci, par exemple en glorifiant les actes terroristes, en sollicitant une personne ou un groupe de personnes en vue de participer aux activités d'un groupe terroriste et en fournissant des instructions sur la manière de mener des attaques, y compris des instructions sur la fabrication d'explosifs. Les contenus diffusés à des fins éducatives, journalistiques, artistiques ou de recherche, ou dans l'objectif de sensibiliser le public de façon à lutter contre les activités terroristes sont protégés en vertu de la proposition de règlement.

À côté de l'obligation de retirer les contenus illicites, le règlement prévoit de multiples garanties pour renforcer la responsabilité et la transparence en ce qui concerne les mesures prises pour retirer les contenus terroristes, ainsi que contre les retraits erronés de contenus relevant de la liberté d'expression légitime. L'article 7 du règlement introduit des obligations de transparence pour les fournisseurs de services d'hébergement. En particulier, ceux-ci sont tenus d'exposer dans leurs conditions générales leur politique en matière de lutte contre la diffusion de contenus

à caractère terroriste. En outre, ils doivent établir des rapports de transparence annuels et y préciser les mesures prises pour identifier et retirer les contenus à caractère terroriste, l'utilisation d'outils automatisés, le nombre de contenus retirés ou rétablis, ainsi que le nombre de plaintes et de procédures de réexamen, avec l'issue qui leur a été réservée.

En décembre 2019, la Commission a adopté une proposition de règlement relatif à la législation sur les services numériques[39] visant à clarifier les responsabilités et à renforcer la responsabilisation des services permettant la diffusion de contenus. La législation sur les services numériques améliore significativement les mécanismes de retrait de contenus illicites ainsi que la protection efficace des droits fondamentaux des utilisateurs en ligne, notamment la liberté d'expression. Elle permet également une surveillance renforcée par les autorités publiques des plateformes en ligne, en particulier celles qui touchent plus de 10 % de la population de l'Union. Les mesures proposées comprennent notamment :

- des mesures visant à lutter contre les marchandises, services ou contenus en ligne qui contreviennent à la loi, dont un mécanisme permettant aux utilisateurs de signaler ces contenus et aux plateformes de coopérer avec des « signaleurs de confiance » ;
- des garanties efficaces pour les utilisateurs, dont la possibilité de contester les décisions relatives à la modération de contenu des plateformes, même lorsque ces décisions se fondent sur les conditions générales d'utilisation des plateformes ;
- des mesures de transparence applicables aux plateformes en ligne pour toutes les décisions de modération de contenu, ainsi que la transparence des algorithmes de recommandation des « très grandes plateformes en ligne » ;
- l'obligation pour les très grandes plateformes en ligne d'empêcher l'utilisation abusive de leurs systèmes en vue de la diffusion de contenu illégal ou la manipulation intentionnelle de leurs services, les plateformes devant prendre des mesures d'atténuation des risques ;
- des audits indépendants des systèmes de gestion des risques, ainsi qu'un accès aux données clés des plus grandes plateformes pour les chercheurs, cet accès ayant pour but de comprendre la façon dont évoluent les risques en ligne ;
- la transparence relative à la publicité destinée aux utilisateurs ainsi que l'obligation pour les très grandes plateformes en ligne de conserver des archives des publicités, avec accès public ;
- une structure de surveillance apte à gérer la complexité de l'espace en ligne : les pays de l'UE joueront à cet égard un rôle de premier plan, avec le soutien d'un nouveau Comité européen des services numériques – pour les très grandes plateformes en ligne, la Commission sera habilitée à exercer une surveillance directe et à prendre des mesures ad hoc.

La proposition de législation passe actuellement à travers les différentes étapes du processus législatif européen et fait l'objet d'un examen attentif par le Conseil de l'Union européenne et le Parlement européen.

France

Le 18 juin 2020, le Conseil constitutionnel français a rendu une décision sur la conformité d'une loi adoptée par le parlement (la « loi Avia ») pour lutter contre la dissémination des contenus haineux sur Internet. Le Conseil constitutionnel a jugé contraires à la Constitution les principales dispositions de la loi, qui auraient réduit le délai dans lequel les fournisseurs de plateformes étaient tenus de réagir aux contenus haineux notifiés (i) à 24 heures pour les contenus haineux signalés par les utilisateurs et (ii) à une heure pour les contenus dangereux spécifiques notifiés

par les autorités françaises comme étant de la pédopornographie ou de la propagande terroriste, sous peine de sanctions pénales.

Le Conseil constitutionnel a jugé ces dispositions inconstitutionnelles au motif qu'elles portaient atteinte à la liberté d'expression sans être « nécessaires, adaptées et proportionnées » à l'objectif poursuivi. Le Conseil constitutionnel a également jugé les délais trop courts, la non-intervention d'un juge français problématique et le risque d'un « excès de censure » ou d'un « excès de blocage ». À la suite de la décision, le Conseil constitutionnel a réduit significativement la portée de la loi Avia. Les dispositions qui subsistent prévoient en substance :

a. une augmentation de l'amende pénale de 75 000 EUR à 250 000 EUR (à multiplier par cinq pour les personnes morales) en cas de non-respect des obligations suivantes (qui existaient déjà dans la législation française) pour les fournisseurs de plateformes en ligne :
    i. l'obligation de saisir et de conserver les données permettant l'identification de quiconque crée un contenu à travers leurs services (pour transmission potentielle aux autorités françaises) ;
    ii. l'obligation de fournir un outil accessible pour notifier le contenu dangereux promouvant les crimes contre l'humanité, provoquant et promouvant les actes de terrorisme, promouvant la haine contre des personnes au titre de leur race, de leur sexe, de leur orientation sexuelle ou de leur identité ou handicap, la pédopornographie, la violence et l'atteinte à la dignité humaine ;
    iii. l'obligation d'informer sans délai les autorités compétentes de tout contenu dangereux cités ci-dessus qui leur a été notifié par des utilisateurs de leurs services ;
    iv. l'obligation de rendre publics les moyens mis en œuvre pour traiter les contenus concernés ;
b. la création d'un parquet spécialisé dans les affaires numériques pour certains types de contenus dangereux répréhensibles sur le plan pénal ;
c. la création d'un « observatoire de la haine en ligne » placé auprès du Conseil supérieur de l'audiovisuel.

À la suite de la décision du Conseil constitutionnel sur la loi Avia, un projet de loi « Respect des principes de la République » a été introduit le 15 janvier 2021 en vue de soumettre les plateformes numériques à certaines obligations concernant la modération des contenus haineux illicites en ligne. La disposition la plus controversée de la loi Avia a été abandonnée, c'est-à-dire celle enjoignant les réseaux sociaux à retirer les contenus haineux manifestement illicites dans les 24 heures. En vertu du projet de loi, les plateformes devront consacrer des « ressources humaines et technologies appropriées » pour modérer les contenus et mettre en place des garanties en matière de procédure, telles que la possibilité pour l'utilisateur d'introduire un recours, en particulier pour les cas les plus graves, comme la résiliation d'un compte. Le projet de loi impose également des obligations accrues en matière de transparence et soumet les plateformes numériques à une surveillance réglementaire stricte. Les exigences spécifiques de ce projet de loi devraient être conformes à celles de la législation sur les services numériques (y compris, notamment, les obligations spécifiques imposées aux très grandes plateformes d'évaluer les risques systémiques présentés par leurs services et d'atténuer ces risques). Le projet de loi devrait expirer à la date d'entrée en vigueur de la législation sur les services numériques, au plus tard fin 2023.

Allemagne

Depuis août 2019, deux projets législatifs ont été introduits – notamment à travers des amendements à la *Netzwerkdurchsetzungsgesetz* (la « NetzDG »), la loi allemande visant à sanctionner les contenus haineux sur les réseaux sociaux, adoptée en 2017 – afin de renforcer

encore le dispositif d'application de la loi envers les réseaux sociaux et, par conséquent, de contribuer à lutter contre les contenus illicites en ligne.

Premièrement, la loi visant à lutter contre l'extrémisme de droite et le crime de haine, adoptée par le parlement allemand le 18 juin 2020, oblige non-seulement les fournisseurs de réseaux sociaux à retirer les contenus illicites, mais, dans les cas particulièrement graves, leur impose de transmettre les contenus concernés à la police judiciaire fédérale allemande. L'intention est de créer les conditions nécessaires pour assurer un retrait le plus rapide possible des contenus concernés et l'engagement de poursuites pénales efficaces. À l'instar de la NetzDG, l'obligation de signalement s'étend également aux infractions pénales afférentes aux contenus terroristes et extrémistes violents en Allemagne, notamment l'infraction de « constitution d'organisations terroristes » (article 129a du code pénal), l'infraction d'« incitation à la haine » (article 130 du code pénal) et l'infraction d'« apologie de la violence » (article 131 du code pénal)[40].

D'autres amendements doivent encore être introduits à travers le projet de loi visant à modifier la loi sanctionnant les contenus haineux sur les réseaux sociaux (NetzDGÄndG), actuellement débattu au parlement. Il serait principalement question de renforcer les droits des utilisateurs en aménageant un droit de réexamen des décisions prises par les fournisseurs de réseaux sociaux sur l'illégalité des contenus et de procéder à des ajustements en relation avec l'affirmation de ces droits sur le plan civil[41].

Irlande

L'Irlande a récemment introduit un projet de loi intitulé « Online Safety and Media Regulation Bill »[42] visant à combler le vide juridique pour lutter contre les contenus dangereux en ligne et établir un cadre réglementaire solide pour faire face à la propagation des contenus dangereux en ligne.

L'expression « contenu dangereux en ligne » comprend les « contenus intégrant des éléments d'incitation à la violence ou à la haine » et les « provocations publiques à commettre une infraction terroriste ».

Le projet de loi exige que la transparence fasse partie intégrante du cadre législatif relatif à la sécurité en ligne et prévoit la nomination d'un commissaire à la sécurité en ligne dont les activités s'exercent au sein d'une commission des médias élargie pour superviser le nouveau cadre réglementaire. Le commissaire mettra ce nouveau cadre réglementaire en action à travers des codes contraignants régissant la sécurité en ligne ainsi qu'en exerçant ses pouvoirs en matière de conformité, d'application des textes et de sanction. Les codes afférents à la sécurité en ligne traiteront d'une large gamme de questions, dont :

- les mesures que les services en ligne sont tenus de prendre pour lutter contre la disponibilité de contenus dangereux en ligne sur leurs services ;
- les plaintes émanant des utilisateurs et/ou les mécanismes de traitement des questions soulevées, actionnés par les services en ligne ;
- les évaluations des risques et des impacts que les services en ligne peuvent effectuer relativement à la disponibilité de contenus dangereux en ligne sur leurs services ;
- les obligations en matière de rapports incombant aux services en ligne.

République de Corée

La Corée a adopté plusieurs lois antiterroristes qui inclue les contenus en ligne. La législation coréenne permet ainsi au responsable d'un organisme public compétent de demander la

coopération du responsable d'une « institution compétente » pour éliminer, suspendre et surveiller les contenus terroristes et extrémistes violents potentiels.

En juillet 2016, l'assemblée générale de l'ONU a adopté une résolution appelant tous les États membres des Nations unies à élaborer un plan d'action national pour prévenir l'extrémisme violent. Le gouvernement de la République de Corée a ainsi mis au point un plan interministériel de prévention de l'extrémisme violent. Ce « plan d'action national pour la prévention de l'extrémisme violent » a été transmis à la commission nationale de lutte contre le terrorisme en janvier 2018 et soumis à l'Organisation des Nations unies. Il comprend des programmes visant à renforcer la coopération public-privé en vue de bâtir un environnement Internet sain et empêcher l'utilisation abusive d'Internet et des technologies de communications par les groupes terroristes.

Le gouvernement coréen participe également à l'initiative Tech Against Terrorism menée par la Direction exécutive du Comité contre le terrorisme des Nations unies (*Counter-Terrorism Executive Directorate*, CTED), qui fait appel aux contributions volontaires pour lutter contre le terrorisme et gérer une plateforme de partage de connaissances à cette fin. Cette plateforme de partage de connaissances sert de centre de partage connecté permettant aux grandes entreprises de transférer leur savoir-faire en matière de lutte contre l'utilisation abusive d'Internet par les groupes extrémistes violents aux petites et moyennes entreprises spécialisées dans les technologies de l'information.

Nouvelle-Zélande

Le gouvernement néo-zélandais continue à faire avancer le projet de loi visant à modifier la classification des films, vidéos et publications aux fins de l'application en urgence de mesures de prévention et de lutte contre les contenus dangereux en ligne (Films, Videos and Publications Classification [Urgent Interim Classification of Publications and Prevention of Online Harm] Amendment Bill)[43]. Le projet de loi a été introduit au parlement le 26 mai 2020 et a fait l'objet d'une première lecture le 10 février 2021. Il prévoit notamment la création d'une infraction pénale pour la diffusion en direct de contenus répréhensibles. La loi correspondante devrait être promulguée avant la fin 2021.

L'infraction pénale relative à la diffusion en direct de contenus répréhensibles s'applique uniquement à l'individu ou au groupe qui diffuse en direct ce contenu. Elle ne concerne pas les hébergeurs de contenus en ligne qui fournissent l'infrastructure en ligne ou la plateforme électronique nécessaire à la diffusion en direct.

Dans le cadre du projet de loi, le responsable de la censure aura les pouvoirs nécessaires pour se prononcer sur le classement provisoire immédiat de toute publication dans des situations où l'apparition soudaine et la distribution virale de contenus répréhensibles sont préjudiciables au bien public. Cette mesure d'évaluation provisoire sera appliquée jusqu'à ce qu'une décision de classement soit prise ou pendant un maximum de 20 jours ouvrables, selon l'événement qui se produit en premier. Le projet de loi autorise également un inspecteur des publications à émettre un avis de retrait pour des contenus répréhensibles en ligne. Ces avis seront adressés à un hébergeur de contenus en ligne et ordonneront le retrait d'un lien spécifique de telle sorte que le contenu concerné ne soit plus visionnable en Nouvelle-Zélande. Un non-respect pourra entraîner une amende pécuniaire au civil.

En outre, le projet de loi clarifie les obligations des hébergeurs de contenus en ligne en relation avec les contenus répréhensibles en vertu de la loi de classement des films, vidéos et publications ainsi que d'autres types de contenus dangereux en ligne entrant dans le champ d'application de la loi de 2015 sur les communications numériques dangereuses (*Harmful Digital Communications Act 2015*,[44] HDCA). La loi HDCA vise à dissuader, prévenir et atténuer les communications numériques dangereuses, et accorde aux victimes de communications numériques un moyen de réparation rapide et efficace. L'article 24 de la loi HDCA dispose qu'en vertu de la législation néo-zélandaise, les hébergeurs de contenus en ligne ne peuvent

être poursuivis pour le fait d'héberger des contenus dangereux sur leurs plateformes s'ils suivent un certain nombre d'étapes lors du dépôt d'une plainte. Le projet de loi indique clairement que lorsque le contenu en ligne concerné est répréhensible, l'article 24 de la loi HDCA ne s'applique pas.

Le ministère de l'Intérieur a récemment mis sur pied un organe réglementaire pour réagir aux signalements de contenus terroristes et extrémistes violents en ligne, cet organe fonctionnant sur le principe d'une coopération volontaire pour retirer ces contenus. Le projet de loi prévoit encore des mécanismes de blocage ou de filtrage des contenus terroristes et extrémistes violents jugés répréhensibles en Nouvelle-Zélande, pour autant que cela devienne nécessaire. Le projet de loi exige que tout filtre de ce type soit sous-tendu par un système de gouvernance et de rapport parfaitement transparent.

Royaume-Uni

Depuis les premières esquisses de son programme visant à réglementer les plateformes et les contenus au Royaume-Uni l'année dernière, le gouvernement a fourni de plus amples informations sur ce à quoi ressemblerait le futur projet de loi sur les dangers en ligne (*Online Harms Bill*) dans sa réponse initiale à la consultation (*Initial Consultation Response*) du 12 février 2020 (DCMS, 2020). Oliver Dowden, le secrétaire d'État du ministère des Affaires numériques, de la culture, des médias et du sport (DCMS), a laissé entendre qu'il envisageait de soumettre le projet de loi sur les dangers en ligne à un examen préalable minutieux, faisant craindre que la législation ne soit pas introduite avant la prochaine session parlementaire (et puisse être repoussée jusqu'en 2022/23).

Le livre blanc sur les dangers en ligne (*Online Harms White Paper*) (Gouvernement du Royaume-Uni, 2019) donne une indication de ce que seront les éléments centraux de la législation en préparation :

- Les services entrant dans le champ d'application de la réglementation devront s'assurer que les contenus illicites sont retirés rapidement et que le risque qu'ils apparaissent est réduit au maximum par la mise en œuvre de systèmes efficaces. Les entreprises seront tenues de « prendre des mesures particulièrement fortes » pour lutter contre les contenus terroristes en ligne ainsi que contre l'exploitation et les abus sexuels des enfants en ligne. (DCMS, 2020)
- Des codes de pratiques volontaires provisoires indiqueront aux entreprises comment traiter les contenus et activités terroristes en ligne. Ces codes visent à permettre au secteur de mettre à niveau ses critères de conformité avant que le régulateur compétent en matière de supervision du nouveau cadre réglementaire (Ofcom) soit fonctionnel.
- Un « système de sanction échelonnée » prévoyant une intensification des mesures de l'amende substantielle au blocage des sites, en passant par la responsabilisation pénale des membres de la direction générale de la plateforme et le blocage du fournisseur de services Internet pour les cas les plus graves.
- Les entreprises devront prouver qu'elles respectent le nouveau principe légal de l'« obligation de vigilance ». L'obligation de vigilance exigera des entreprises qu'elles assument une plus grande responsabilité pour les contenus et comportements dangereux apparaissant sur leurs plateformes. Elles devront s'assurer qu'elles ont mis en place des systèmes et des processus efficaces pour réduire et réagir aux dangers en ligne. Un régulateur indépendant aura pour mission de superviser le respect de cette obligation de vigilance. Le projet de loi *Online Harms* propose de contraindre les entreprises technologiques à faire respecter leur obligation de vigilance comme suit :
  - Les conditions d'utilisation doivent être mises à jour de façon à ce qu'elles mentionnent explicitement les contenus appropriés (ou non) sur leurs plateformes ;
  - Des rapports de transparence annuels doivent être établis et publiés ;
  - Un système de gestion des plaintes des utilisateurs facile d'accès doit être mis en place ;

- Les entreprises doivent être soumises à l'obligation de réagir aux plaintes des utilisateurs dans un « délai approprié » à définir par l'Ofcom.

États-Unis

L'approche adoptée par les États-Unis face aux contenus terroristes et extrémistes violents en ligne est principalement inspirée par le premier amendement à la Constitution états-unienne qui dispose que « le Congrès n'adoptera aucune loi […] pour limiter la liberté d'expression ». D'une manière générale, le premier amendement protège une liberté d'expression au sens large, même les propos odieux ou insultants, et interdit toute restriction préalable ou censure de la parole par le gouvernement. Le gouvernement peut toutefois interdire les discours visant à produire ou à inciter à la commission d'actions illégales imminentes ou qui seraient susceptibles d'avoir un tel effet. Ainsi, plutôt que de criminaliser les propos haineux ou odieux et les discours incitant à la violence ou défendant des causes ou des groupes dangereux, les États-Unis se sont concentrés sur la poursuite des activités criminelles prônant la violence ainsi que sur la promotion de discours alternatifs crédibles, en en faisant leur principal moyen pour saper et contrer le message terroriste.

Plusieurs lois des États-Unis criminalisent les comportements verbaux faisant l'apologie des actions violentes, y compris les actes terroristes. Ainsi, l'article 373 du titre 18 du Code des États-Unis érige en infraction le fait de pousser ou d'ordonner à une autre personne de commettre un crime contre une autre personne ou des biens en violation des lois des États-Unis, ou d'inciter une autre personne à le faire, en menaçant, en tentant d'utiliser ou en utilisant effectivement la force physique.

En outre, les dispositions relatives à l'apport d'un soutien matériel aux organisations terroristes étrangères de l'article 2339B de ce même titre 18 du Code américain s'appliquent aux actes commis en connaissance de cause en vue de soutenir des organisations terroristes désignées comme telles, ou sous la direction de ou en coordination avec de telles organisations, et dont l'auteur sait qu'il s'agit d'organisations terroristes.

En vertu de la législation américaine, les fournisseurs de services en ligne sont généralement exempts de toute responsabilité quant aux propos de leurs utilisateurs et sont dégagés de toute responsabilité en ce qui concerne leurs décisions de modération des contenus, sauf dans des circonstances limitées, notamment en cas d'infraction au droit pénal fédéral (voir l'article 230 de la loi américaine de 1934 sur les communications). Le cadre législatif régissant le principe de la responsabilité intermédiaire aux États-Unis donne aux fournisseurs de services en ligne la capacité de modérer l'utilisation de leurs plateformes pour certains types de discours qui ne pourraient pas être interdits par le gouvernement.

Annexe A – Liste des 50 services les plus prisés

Rang	Nom du service (société mère)	Nombre d'utilisateurs actifs, de comptes d'utilisateur ou de visiteurs uniques (millions)	Type de service	Publie des rapports de transparence	A réagi ou transmis des commentaires concernant son profil
1	Facebook (Facebook, Inc.)	2 603 (juillet 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Plateforme de réseau social et de diffusion de flux vidéos	O	O
2	YouTube (Alphabet, Inc.)	2 000 (juillet 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Plateforme de diffusion de flux vidéos	O	O
3	WhatsApp (Facebook, Inc.)	2 000 (juillet 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Application de messagerie	N	O
4	Facebook Messenger (Facebook, Inc.)	1 300 (juillet 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Application de messagerie	N	O
5	iMessage/FaceTime (Apple, Inc)	1 300 (janvier 2019) (Elmer-Dewitt, 2019)	Applications de messagerie et de bavardage vidéo	N	N

6	Weixin/WeChat (Tencent Holdings Ltd)	1 203 (juillet 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Plateforme de réseau social, de partage de contenus et de messagerie	N	N
7	Instagram (Facebook, Inc.)	1 082 (juillet 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Plateforme de réseau social	O	O
8	Tik Tok (ByteDance Technology Co.)	800 (juillet 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Application de vidéos courtes	O	N
9	QQ (Tencent Holdings Ltd)	694 (juillet 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Site de messagerie instantanée et portail	N	N
10	Youku Tudou (Alibaba Group Holding Limited)	580 (août 2019) (Youku Tudou Inc. (NYSE: YOKU), n.d.)	Plateforme de diffusion de flux vidéos (d'utilisateurs et de médias)	N	N
11	Weibo (Sina Corp.)	550 (juillet 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Plateforme de réseau social	N	N
12	QZone (Tencent Holdings Ltd)	517 (juillet 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Plateforme de réseau social	N	N
13	iQIYI (Baidu, Inc.)	476 (décembre 2019) (Statista, 2019)	Plateforme de diffusion de flux vidéos (d'utilisateurs et de médias)	N	N

14	Reddit (Reddit, Inc.)	430 (juillet 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Agrégateur d'actualités de réseaux sociaux, site internet de classement de contenus et de discussion	O	O
15	Kuaishou (Beijing Kuaishou Technology Co., Ltd)	400 (juillet 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Application de vidéos courtes	N	N
16	Telegram (Telegram Messenger LLP)	400 (avril 2020) (Singh, 2020)	Application de messagerie	N	N
17	Snapchat (Snap, Inc.)	397 (juillet 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Plateforme de réseau social	N	O
18	Pinterest (Pinterest, Inc.)	367 (juillet 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Plateforme de réseau social	N	N
19	Twitter (Twitter, Inc.)	326 (juillet 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Plateforme de réseau social spécialisée dans les messages brefs	O	N
20	Douban (Information Technology Company, Inc.)	320 (juillet 2019) (Kemp, Digital 2019: Q3 Global Digital Statshot, 2019)	Plateforme de réseau social	N	N
21	LinkedIn (Microsoft, Inc.)	310 (juillet 2019) (Kemp, Digital 2019: Q3 Global Digital Statshot, 2019)	Plateforme de réseau social professionnel	N	O

22	Baidu Tieba (Baidu, Inc.)	300 (mars 2020) (Marketing to China, 2020)	Plateforme de communications en ligne	N	N
23	Skype (Microsoft, Inc.)	300 (juin 2019) (Perez, Skype publicly launches screen sharing on iOS and Android, 2019)	Application de bavardage vidéo et d'appels vocaux	O	O
24	Quora (Quora, Inc.)	300 (septembre 2018) (Marketing Land, 2018)	Site de questions-réponses	N	N
25	Xigua (ByteDance Technology Co.)	270 (décembre 2019) (Chen, 2020)	Application de diffusion de vidéos courtes	N	N
26	Viber (Rakuten, Inc.)	260 (juillet 2019) (Kemp, Digital 2019: Q3 Global Digital Statshot, 2019)	Application de messagerie	N	O
27	Discord (Discord, Inc.)	250 (juillet 2019) (Kemp, Digital 2019: Q3 Global Digital Statshot, 2019)	Plateforme de bavardage	O	N
28	Vimeo (Vimeo, Inc.)	240 (septembre 2018) (Bicknell, 2018)	Application de diffusion de flux vidéos	N	N
29	IMO (PageBites, Inc.)	211 (avril 2019) (YY Inc. - IR Site, 2019)	Application de bavardage vidéo et d'appels vocaux	N	N
30	LINE (Line Corporation)	194 (janvier 2019) (Kemp, Digital 2019: Global Digitial Overview, 2019)	Application de messagerie	N	N
31	Huoshan (ByteDance Technology Co.)	170 (décembre 2019) (Chen, 2020)	Application de diffusion de vidéos courtes	N	N

32	Ask.fm (IAC [InterActiveCorp])	160 (avril 2020) (Kallas, 2020)	Plateforme de réseau social	N	O
33	YY Live/Huya (YY, Inc.)	157 (novembre 2019) (Yahoo! Finance, 2019)	Plateforme de diffusion en direct	N	N
34	Twitch (Amazon.com, Inc.)	140 (juillet 2020) (Iqbal, 2020)	Plateforme de diffusion en direct	O	O
35	Tumblr (Automattic, Inc.).	115 (avril 2020) (Kallas, 2020)	Plateforme de microblogage et de réseau social	N	N
36	Flickr (SmugMug, Inc.)	112 (avril 2020) (Kallas, 2020)	Services d'hébergement d'images et de vidéos	N	N
37	VK (Mail.Ru Group)	97 (avril 2020) (Kallas, 2020)	Plateforme de réseau social	N	O
38	Medium (A Medium Corporation)	86 (août 2018) (Wickey, 2018)	Plateforme de publication en ligne	N	O
39	Odnoklassniki (Mail.Ru Group)	71 (avril 2020) (Kallas, 2020)	Plateforme de réseau social	N	N
41	Haokan (Baidu, Inc.)	69 (juin 2019) (Chen, 2020)	Application de diffusion de vidéos courtes	N	N
41	Smule (Smule, Inc.)	52 (juillet 2018) (Solsman, 2018)	Plateforme de partage de musique et de vidéos d'utilisateurs	N	N
42	KaKao Talk (Daum Kakao Corporation)	50 (juin 2019) (Statista, 2019)	Application de messagerie	N	O
43	Deviantart (DeviantArt, Inc.)	45 (2016) (DeviantArt Media Kit, n.d.)	Plateforme d'iconographie, de vidéographie et de photographie en ligne	N	N
44	Meetup (WeWork Companies, Inc.)	35 (avril 2020) (Kallas, 2020)	Plateforme de réseau social internet	N	N
45	4chan (4chan Community Support LLC)	22 (août 2019) (4chan, n.d.)	Plateforme de partage de contenus	N	N

Les données relatives au nombre d'utilisateurs actifs mensuels (*Monthly Active User*, MAU) ne sont pas disponibles pour tous les services de partage de contenus en ligne que les terroristes et extrémistes violents ont utilisés. Toutefois, les indicateurs disponibles laissent entendre que les services pour lesquels ces données n'existent pas devraient être intégrés dans la liste des 50 premiers. Le tableau qui suit constitue donc la continuation du précédent avec cinq services supplémentaires, mais sans classement (rang), car des indicateurs autres que le nombre d'utilisateurs actifs mensuels révèlent qu'ils sont pertinents. Toutefois, il n'était pas possible d'établir une comparaison complète avec les services ci-dessus. Quoi qu'il en soit, aux fins du présent rapport, la composition générale du groupe des 50 premiers revêt plus d'importance que le classement individuel des services concernés.

Nom du service (société mère)	**Part de marché globale indicative**	**Type de marché/service**	**Rapport de transparence sur les contenus terroristes et extrémistes violents**	**A réagi ou transmis des commentaires concernant son profil**
Google Drive (Alphabet, Inc.)	34,35 % (octobre 2019) (Datanyze, 2020)	Partage infonuagique de fichiers	N	O
Dropbox (Dropbox, Inc.)	21,23 % (octobre 2019) (Datanyze, 2020)	Partage infonuagique de fichiers	N	O
Microsoft OneDrive (Microsoft, Inc.)	12,07 % (octobre 2019) (Datanyze, 2020)	Partage infonuagique de fichiers	O	O

Nom du service (société mère)	**Part de marché globale indicative ou nombre mensuel de pages consultées**	**Type de marché/service**	**Rapport de transparence sur les contenus terroristes et extrémistes violents**	**A réagi ou transmis des commentaires concernant son profil**
Wordpress.com (Automattic, Inc.)	60% (avril 2019) (Kinsta, 2011-2019)	Système de gestion de contenus	O	O
Wikipedia (Wikimedia Foundation)	18 milliards de pages consultées par mois (janvier 2016) (Pew Research Center, 2016) ; 10^{e} site internet le plus visité à l'échelle mondiale (Alexa, 2019)	Encyclopédie en ligne	N	N

Annexe B – Profils des 50 services les plus prisés

1. Facebook[45]

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition particulière des contenus terroristes et extrémistes violents. Facebook est toutefois l'un des rares services internet à disposer d'une définition bien précise du terrorisme et des termes associés. Dans la section des « standards de la communauté Facebook » intitulée « Individus et organisations dangereux » (Facebook), Facebook indique que les organisations ou individus qui revendiquent des objectifs violents ou qui sont impliqués dans des activités violentes ne sont pas les bienvenus sur Facebook. Ces organisations et individus sont définis comme celles et ceux qui sont impliqués dans les activités suivantes : ● les activités terroristes ; ● la haine organisée ; ● les tueries (y compris les tentatives) ou homicides multiples ; ● la traite des personnes ; ● la violence ou les activités criminelles organisées. Les contenus soutenant ou faisant l'apologie de groupes, dirigeants ou individus impliqués dans ces activités sont supprimés. Les individus (vivants ou décédés) et les groupes suivants ne peuvent pas maintenir une présence (par exemple, en possédant un compte, une page ou un groupe) sur la plateforme : organisations terroristes, terroristes, organisations animées par la haine (ainsi que leurs dirigeants et leurs membres prééminents) et les meurtriers de masse et de plusieurs homicides. Les organisations terroristes et les terroristes comprennent les acteurs non étatiques qui : ● prennent part à des actes de violence intentionnels et prémédités, les défendent ou les soutiennent de manière active ; ● causent ou tentent de causer la mort, des blessures ou des dommages graves à des civils ou à toute autre personne ne prenant pas directement part aux hostilités dans le cadre d'un conflit armé, ou des dommages

	matériels graves associés à la mort, à des blessures graves ou à des dommages graves dont les victimes sont des civils ; ● dans le but d'assujettir, d'intimider ou d'influencer une population civile, un gouvernement ou une organisation internationale ; ● pour atteindre un objectif politique, religieux ou idéologique. Les organisations incitant à la haine sont définies comme des associations de trois personnes ou plus qui sont organisées sous un nom, un signe ou un symbole et dont l'idéologie, les déclarations ou les actions physiques portent atteinte à des individus en fonction de caractéristiques, notamment l'origine ethnique, l'affiliation religieuse, la nationalité, le genre, le sexe, l'orientation sexuelle, une maladie grave ou le handicap. Un homicide est considéré comme une tuerie s'il fait au moins trois victimes en un incident. Est considéré comme meurtrier en série tout individu ayant commis au moins deux meurtres au cours de plusieurs incidents ou à plusieurs endroits. Facebook interdit les symboles qui représentent l'une des organisations ou l'un des individus cités ci-dessus s'ils ne sont pas partagés avec un contexte condamnant le contenu ou en discutant de façon neutre. Les contenus prônant l'une des organisations ou l'un des individus cités ci-dessus ou tout acte commis par eux sont interdits. De même, Facebook ne permet pas la conduite d'activités de coordination menées en vue de soutenir les organisations ou individus visés ci-dessus ou leurs actes. En outre, Facebook interdit les contenus qui représentent ou défendent d'une quelconque façon des événements qu'il désigne comme étant des attaques terroristes, des incitations à la haine ou des tueries. Enfin, à la section intitulée « Violence et provocation » des standards de la communauté Facebook (Facebook), Facebook indique qu'il supprime les propos qui encouragent ou permettent des violences graves. En particulier, les utilisateurs ne sont pas autorisés à publier : ● des menaces pouvant entraîner la mort (ou toute autre forme de violence très grave) d'une cible, les menaces correspondant à l'un des éléments suivants : o déclaration de l'intention de commettre un acte de violence très grave, o appel à commettre un acte de violence très grave, y compris les contenus n'indiquant pas de cible précise, mais un symbole représentant la cible ou l'image d'une arme pour représenter la violence, o déclaration prônant un acte de violence très grave, o déclaration intentionnelle ou conditionnelle de commettre un acte de violence très grave ;

	● des contenus qui demandent ou proposent des services de location d'un tueur (par exemple, tueur à gages, mercenaire, assassin) ou préconisent de faire appel à un tueur à gages, un mercenaire ou un assassin contre une cible ; ● des reconnaissances, des déclarations d'intention ou des incitations, des appels à action ou des déclarations intentionnelles ou conditionnelles à enlever une cible ; ● des menaces pouvant causer des blessures graves (violence de gravité moyenne) à des personnes, des personnalités publiques mineures, des personnes ou des groupes vulnérables, les menaces correspondant à l'un des éléments suivants : o déclaration de l'intention de commettre un acte de violence, o déclaration faisant l'apologie de la violence, o appel à commettre un acte de violence de gravité moyenne, y compris les contenus n'indiquant pas de cible précise, mais un symbole représentant la cible, o déclaration intentionnelle ou conditionnelle de commettre un acte de violence, o contenu concernant d'autres cibles que des personnes, des personnalités publiques mineures, des personnes ou des groupes vulnérables, et tout élément indiqué ci-après considéré comme crédible : ▪ déclaration de l'intention de commettre un acte de violence, ▪ appel à commettre un acte de violence, ▪ déclaration faisant l'apologie de la violence, ▪ déclaration intentionnelle ou conditionnelle de commettre un acte de violence ; ● des menaces pouvant causer un préjudice physique (ou toute autre forme de violence de faible gravité) à des personnes (autosignalement obligatoire) ou des personnalités publiques mineures, les menaces correspondant à l'un des éléments suivants : o déclaration d'intention, o appel à action, o déclaration intentionnelle ou conditionnelle de commettre un acte de violence de faible gravité ou incitant à le faire ; ● des représentations de personnes ou de personnalités publiques mineures qui ont été manipulées pour contenir des menaces de violence sous forme de texte ou d'image (cible, flèche, pistolet pointé sur la tête, etc.) ;

	● des contenus créés dans le but exprès de désigner une personne comme un membre appartenant à un groupe à risque précis et reconnaissable ; ● des instructions sur la façon de fabriquer ou d'utiliser des armes s'il existe des preuves de l'intention de blesser gravement ou de tuer des personnes, telles que : o des messages énonçant clairement cette intention, o des photos ou des vidéos montrant ou simulant le résultat (blessure grave ou mort) dans le cadre des instructions, o sauf si les contenus cités ci-dessus sont partagés dans un but d'autodéfense récréative, à des fins d'entraînement militaire, pour des jeux vidéo commerciaux ou pour relater des événements d'actualité (publiés sur une page ou avec un logo de nouvelles ou d'actualités) ; ● des instructions sur la façon de fabriquer ou d'utiliser des explosifs, sauf si le contexte indique clairement que le contenu est publié à des fins non violentes (par exemple, jeux vidéo commerciaux, but scientifique/pédagogique clair, feux d'artifice ou matériel de pêche) ; ● des contenus comportant des déclarations d'intention, des appels au passage à l'acte ou faisant l'apologie de la violence de forte ou moyenne gravité en raison de la tenue d'élections, de l'inscription d'électeurs ou des résultats d'une élection ; ● de fausses informations contribuant à provoquer de la violence imminente ou des préjudices physiques ; ● des appels à action, des déclarations de l'intention d'apporter des armes dans un lieu donné, tel qu'un lieu de culte, ou des encouragements à le faire.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.facebook.com/communitystandards/.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Oui, elles sont disponibles sur https://about.fb.com/news/2019/05/protecting-live-from-abuse/. En particulier, Facebook applique une règle dite du « premier avertissement » aux contenus interdits diffusés en direct. Selon cette règle, quiconque enfreint les politiques les plus importantes de Facebook se voit bloqué dans l'utilisation de la diffusion en direct pendant une période définie, par exemple 30 jours, à partir du jour de la première infraction.
4. Politiques et procédures de mise en œuvre et d'application	Lorsqu'un contenu enfreint ses normes communautaires (les « standards de la communauté »), Facebook le retire de la plateforme.

des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	
4.1 Notifications des suppressions ou des autres décisions de sanction	La personne qui a publié le contenu est avertie de la suppression lorsque celle-ci a été effectuée et a la possibilité de demander une révision ou d'accepter la décision (Facebook).
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Si l'utilisateur demande une révision, le contenu est soumis à un nouvel examen. Tant que la procédure de réexamen est en cours, le contenu reste invisible et inaccessible aux autres utilisateurs de Facebook. Les évaluateurs ne savent pas que la publication a déjà fait l'objet d'un premier examen. Autant qu'il soit possible d'en juger d'après la formulation du document « Understanding the Community Standards Report » (Facebook), l'examen serait le fait d'une seule et unique personne. Si l'évaluateur est d'accord avec la décision initiale, le contenu n'est pas republié sur Facebook. En revanche, s'il n'est pas d'accord avec la décision initiale et estime que le contenu n'aurait pas dû être supprimé, celui-ci sera examiné par un troisième évaluateur. C'est la décision de ce dernier qui déterminera si le contenu est autorisé ou non sur Facebook.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Facebook détecte les violations de ses règles, y compris celle relative à la propagande terroriste, par la mise en œuvre simultanée de plusieurs moyens : outils techniques, rapports des utilisateurs et examens par ses équipes (Facebook). En particulier, Facebook recourt à l'intelligence artificielle pour lutter contre le terrorisme, celle-ci impliquant des techniques telles que la mise en correspondance d'images, l'analyse sémantique du discours, la suppression de foyers d'activités terroristes et la collaboration interplateforme avec d'autres entités détenues par Facebook (c'est-à-dire WhatsApp et Instagram). Il y a trois ans environ, Facebook a commencé à recourir à des outils d'apprentissage automatique pour analyser les publications sur Facebook qui pourraient faire transparaître des soutiens à l'EIIL ou à Al-Qaida (Facebook, 2018). Depuis, Facebook a généralisé le recours à ces techniques pour détecter et retirer les contenus liés à d'autres groupes terroristes et organisations incitant à la haine. Facebook est à présent en mesure de détecter les textes intégrés dans des images et des

	vidéos pour en comprendre le contexte global et a mis au point des techniques de mise en correspondance de médias pour identifier les contenus identiques ou fort semblables à des photos, vidéos, textes et même pistes sonores que Facebook a déjà retirés. Lorsque Facebook s'est lancé dans la détection des organisations incitant à la haine, il s'est concentré sur les groupes qui constituaient à l'époque la plus grande menace de perpétrer des actes de violence. Il a ensuite élargi son champ d'action de façon à détecter un plus grand nombre de groupes liés à différentes idéologies fondées sur la haine et génératrices d'extrémismes violents dans différentes langues. Outre la mise au point de nouveaux outils, Facebook a adapté les stratégies issues de ses efforts de lutte contre le terrorisme, celles-ci consistant notamment à tenir compte de signaux extérieurs à la plateforme pour identifier les contenus dangereux publiés en ligne et à mettre en œuvre des procédures d'audit pour déterminer la pertinence des décisions de ses systèmes d'intelligence artificielle sur le temps (Facebook, 2020). Facebook a rapporté que la vidéo de l'attaque de Christchurch n'a pas déclenché ses systèmes de détection automatiques parce que le service ne disposait pas de suffisamment de contenus représentant des séquences d'événements violents filmés à la première personne pour entraîner de manière efficace sa technologie d'apprentissage automatique. À la suite de ces événements, Facebook s'est mis à collaborer avec les autorités publiques et les services chargés de l'application des lois des États-Unis et du Royaume-Uni pour obtenir des séquences filmées provenant de leurs propres programmes d'entraînement à l'utilisation d'armes à feu, constituant ainsi une source précieuse de données pour entraîner ses systèmes. Avec cette initiative, Facebook cherche à améliorer la détection de séquences vidéos d'événements violents réels filmés à la première personne tout en évitant de signaler par erreur d'autres types de vidéos, tels que les contenus fictionnels de films et de jeux vidéo. (Facebook, 2019) Facebook fait observer que l'intelligence artificielle n'est pas en mesure de tout prendre. Les rapports des utilisateurs jouent ainsi un rôle fondamental dans la détection de contenus répréhensibles et permettent à Facebook non seulement d'identifier les nouveaux contenus problématiques rapidement, mais aussi d'améliorer les signaux mis en œuvre dans ses techniques de détection des infractions à ses règles (Facebook). Enfin, Facebook a mis en place une équipe spéciale baptisée « *Community Operations Team* », qui examine les contenus dans le cadre de leur contexte élargi pour déterminer s'ils enfreignent ses règles. Cette équipe compte parmi ses membres des spécialistes du terrorisme. Elle passe en revue les rapports 24 heures sur 24, 7 jours sur 7, la toute grande majorité des rapports étant examinés dans les 24 heures (Facebook). Selon Facebook, qu'il y ait eu identification par ses techniques ou rapport des utilisateurs, une violation potentielle signalée se traduit par un rapport dans son système. Facebook donne la

	priorité aux rapports ayant des implications sur le plan de la sécurité, comme ceux portant sur les contenus liés au terrorisme ou au suicide. Facebook applique ensuite des procédés techniques, une procédure de contrôle humain ou une combinaison des deux pour déterminer si un élément de contenu spécifique enfreint ses règles. Si le contenu est transmis à l'équipe de contrôle humain, les membres de l'équipe appliquent les règles de Facebook ainsi qu'un processus par étapes pour prendre la décision la plus adaptée et la plus cohérente compte tenu du type de violation. Facebook fournit aussi à ses modérateurs des outils pour analyser les contenus signalés ainsi que le contexte disponible, ce dernier étant indispensable pour identifier le problème et pour déterminer si un élément de contenu viole effectivement une norme communautaire (Facebook). Le coût marginal du recours à des outils d'intelligence artificielle pour identifier les contenus terroristes et extrémistes violents est probablement très bas (même si les coûts fixes peuvent être importants), tandis que le coût marginal du recours à cette fin à des modérateurs humains est probablement relativement élevé. Facebook est un membre fondateur du GIFCT et participe au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	Les conséquences d'une infraction aux normes communautaires de Facebook varient en fonction de la gravité de la violation et de l'historique attaché à la personne sur la plateforme. Les contenus interdits peuvent être retirés. Au-delà, Facebook peut envoyer un avertissement après une première infraction, mais peut ensuite limiter les possibilités de publication de l'utilisateur sur Facebook ou désactiver son profil si celui-ci continue à enfreindre les règles de la plateforme. Facebook peut également avertir les services chargés de l'application des lois s'il estime qu'il existe un réel risque d'atteinte à l'intégrité corporelle de quelqu'un ou une menace directe pour la sécurité publique.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Oui (Facebook, 2017-2020). Facebook publie des rapports de transparence sur l'application de ses règles collectives (les « standards de la communauté »), dont une section s'intéresse particulièrement aux organisations dangereuses actives dans le terrorisme et la haine organisée, tandis qu'une autre traite des contenus violents et explicites. Précisons que Facebook indique ne pas tolérer les contenus qui font l'éloge d'individus ou de groupes engagés dans des activités terroristes ou incitant à la haine, les cautionnent ou les représentent. Facebook applique cette règle pour les activités et les groupes terroristes à l'échelle régionale et mondiale. Depuis novembre 2019, la partie consacrée à la propagande terroriste du rapport mesure les actions menées contre toutes les organisations terroristes et non plus uniquement contre la propagande associée à l'EIIL, Al-Qaida et leurs groupes affiliés (Facebook, 2020).

8. Informations ou types de données figurant dans les rapports de transparence	Dans son dernier rapport, daté de novembre 2020, Facebook faisait figurer les cinq champs suivants tant dans la section « Dangerous Organizations: Terrorism and Organized Hate » que dans celle intitulée « Violent and Graphic Content » : - *Incidence (incidence des violations liées aux contenus terroristes et aux contenus violents et explicites sur Facebook)* L'indicateur d'incidence correspond au pourcentage de vues associées à des contenus liés au terrorisme et à des contenus violents et explicites. Ainsi, la limite supérieure estimée par Facebook est de 0,05 % de vues de contenus enfreignant ses règles en matière de terrorisme au deuxième trimestre de 2020. En d'autres termes, sur 10 000 vues entrant dans le champ d'application de la politique de Facebook relative au terrorisme, au maximum 5 présentaient des contenus en infraction. (Ces chiffres sont ceux après détermination définitive du statut des contenus, sans tenir compte des contenus initialement signalés comme enfreignant potentiellement la règle, mais jugés admissibles ultérieurement.) - *Contenus pour lesquels des mesures ont été prises (nombre de contenus pour lesquels des mesures ont été prises par Facebook)* Facebook indique qu'un élément de contenu peut correspondre à « un nombre quelconque d'éléments » (Facebook), y compris une publication, une photo, une vidéo ou un commentaire. La prise de mesures peut inclure le retrait d'un élément de contenu de Facebook, le recouvrement de photos ou de vidéos qui pourraient être dérangeantes pour certains publics par un avertissement ou la désactivation de comptes. Les contenus pour lesquels des mesures ont été prises désignent le nombre total d'éléments de contenu pour lesquels Facebook a pris des mesures pendant une période de rapport donnée au titre qu'ils enfreignaient ses standards de la communauté (dans le cas présent, les règles relatives aux contenus relevant du terrorisme ainsi qu'aux contenus violents et explicites). - *Taux de proactivité (pourcentage de contenus en infraction pour lesquels des mesures ont été prises par Facebook avant que les utilisateurs les signalent)* Cet indicateur présente le pourcentage de contenus détectés et pour lesquels des mesures ont été prises par Facebook en raison de leur lien avec des organisations dangereuses ou en raison de leur nature violente et explicite avant que les utilisateurs les signalent. Il comptabilise les détections opérées à la fois par les outils d'intelligence artificielle et par les évaluateurs humains de Facebook. - *Recours (nombre de recours contre des décisions prises par Facebook à l'encontre de contenus en infraction)* Cet indicateur indique le nombre total de contenus pour

	lesquels des mesures ont été prises et qui ont fait l'objet d'une demande de réexamen sur la période couverte par le rapport. - *Contenus rétablis (nombre de contenus que Facebook a rétablis après les avoir retirés)* Le nombre de contenus rétablis correspond au nombre d'éléments de contenu que Facebook a rétablis au cours de la période considérée après avoir pris des mesures à leur encontre. Facebook fait également état des tendances récentes concernant les contenus relevant de la haine organisée et du terrorisme pour lesquels des mesures ont été prises. Ainsi, son dernier rapport de transparence note que les contenus relevant de la haine organisée pour lesquels des mesures ont été prises ont diminué de 4,7 millions d'éléments de contenu au premier trimestre de 2020 à 4 millions au deuxième trimestre de la même année, tandis que les contenus pour lesquels des mesures ont été prises au titre de la lutte contre le terrorisme ont augmenté de 6,3 millions au premier trimestre de 2020 à 8,7 millions au deuxième trimestre.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	- *Incidence.* Ce chiffre indique le nombre estimé de vues d'un contenu enfreignant les règles divisé par le nombre estimé de vues de la totalité des contenus sur Facebook sur la période couverte par le rapport. Par exemple, une incidence d'organisations dangereuses de 0,18 % à 0,20 % signifie que sur 10 000 vues de contenus, 18 à 20, en moyenne, correspondaient à des vues de contenus enfreignant les normes de Facebook concernant les organisations dangereuses. L'incidence fournit une indication de la fréquence à laquelle est vu un contenu interdit, et non le nombre total de ce type de contenus. Elle est estimée à partir d'échantillons de contenus prélevés dans différentes rubriques de Facebook, telles que les groupes ou le fil d'actualité. Pour les violations relevant du terrorisme, en particulier, Facebook fournit uniquement une estimation de la limite supérieure, parce qu'il se dit « convaincu que l'incidence des vues des contenus enfreignant ces règles est inférieure à cette limite » (Facebook). Facebook détaille la méthodologie adoptée en ce qui concerne le calcul de l'incidence dans son article consacré à la mesure de l'incidence des contenus en infraction publiés sur Facebook, « Measuring Prevalence of Violating Content on Facebook » (Facebook, 2019). - *Contenus pour lesquels des mesures ont été prises.* Les contenus pour lesquels des mesures ont été prises désignent le nombre total d'éléments de contenu pour lesquels Facebook a pris des mesures pendant une période de rapport donnée au titre qu'ils enfreignaient ses règles applicables aux contenus. Facebook ne compte pas les cas de figure dans lesquels l'affaire est remontée à un service chargé de l'application des lois. Cet indicateur comprend tant les contenus pour lesquels

	Facebook a pris des mesures après signalement par un utilisateur que ceux que Facebook a repérés de manière proactive. Il couvre les contenus de Facebook et de Messenger. - *Taux de proactivité.* Cet indicateur est le résultat de la division du nombre d'éléments de contenu pour lesquels des mesures ont été prises à la suite d'une détection par Facebook avant que les utilisateurs les signalent par le nombre total d'éléments de contenu pour lesquels des mesures ont été prises. Il couvre les contenus de Facebook et de Messenger. - *Nombre de recours.* Cet indicateur indique le nombre total de contenus pour lesquels des mesures ont été prises et qui ont fait l'objet d'une demande de réexamen sur la période couverte par le rapport. Il couvre les contenus de Facebook et de Messenger. Facebook fait observer que cet indicateur mesure le nombre d'éléments de contenu ayant fait l'objet d'un recours dans le trimestre, le nombre de contenus rétablis pendant le trimestre faisant quant à lui l'objet d'un autre indicateur (voir ci-dessous). Comme certains éléments de contenu peuvent avoir été rétablis au trimestre suivant et que certains contenus rétablis peuvent avoir fait l'objet d'un recours au trimestre précédent, une comparaison directe de ces indicateurs n'est pas possible. - *Contenus rétablis.* Pour calculer cet indicateur, Facebook compte le nombre d'éléments de contenu rétablis au cours de la période considérée après avoir pris des mesures à leur encontre. Facebook peut rétablir des contenus soit dans le cadre d'un recours introduit à la suite d'une décision de retrait, soit lorsqu'une raison de le faire parvient à la connaissance de Facebook. Cet indicateur couvre seulement les contenus de Facebook.
10. Fréquence de publication des rapports de transparence	Facebook publie des rapports de transparence tous les trimestres depuis août 2020. Son dernier rapport porte sur le deuxième trimestre de 2020. Actuellement, les données disponibles concernent la période allant du quatrième trimestre de 2017 au deuxième trimestre de 2020.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. Voir les sections 7 à 9 ci-dessus.

2. YouTube

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition particulière des contenus terroristes et extrémistes violents. Toutefois, le règlement de la communauté de YouTube comprend différentes explications concernant notamment les contenus terroristes et extrémistes violents. Les règles concernant les organisations criminelles violentes, par exemple, spécifient que les contenus visant à faire l'éloge ou la promotion des organisations criminelles violentes ainsi que ceux visant à les soutenir ne sont pas autorisés sur YouTube. En outre, ces organisations sont interdites sur YouTube à quelque fin que ce soit, y compris le recrutement. Le règlement ne répertorie cependant pas ces organisations et ne renvoie pas non plus à une liste extérieure. Quoi qu'il en soit, le règlement interdit les types de contenus suivants : • Contenus produits par des organisations criminelles ou terroristes violentes • Contenus faisant l'apologie ou commémorant des personnalités terroristes ou criminelles dans le but d'encourager d'autres personnes à commettre des actes de violence • Contenus louant ou justifiant des actes de violence commis par des organisations criminelles ou terroristes violentes • Contenus destinés à recruter de nouveaux membres pour des organisations criminelles ou terroristes violentes • Contenus présentant des otages, ou mis en ligne dans le but de solliciter, de menacer ou d'intimider des personnes au nom d'une organisation terroriste ou criminelle violente • Contenus exposant les insignes, les logos ou les symboles d'organisations criminelles ou terroristes violentes afin d'en faire l'apologie ou de les promouvoir Si des contenus liés au terrorisme ou au milieu criminel sont publiés à des fins éducatives, documentaires, scientifiques ou artistiques, il faudra fournir suffisamment d'informations au sein même de la vidéo ou de l'audio pour que les spectateurs puissent en comprendre le contexte. Les règles concernant les organisations criminelles violentes fournissent également différents exemples de contenus non autorisés sur YouTube, comme suit : • Remise en ligne brute et non modifiée de contenus créés par des organisations terroristes ou criminelles • Apologie de chefs terroristes ou de leurs crimes à travers des chansons ou des commémorations • Apologie d'organisations terroristes ou criminelles à travers des chansons ou des commémorations • Contenu orientant les utilisateurs vers des sites prônant une idéologie terroriste, diffusant du contenu interdit ou dont le but est de recruter de nouveaux membres • Contenus de jeux vidéo développés ou modifiés pour faire l'apologie d'un événement violent ou de ses auteurs, ou pour soutenir des organisations terroristes ou criminelles violentes De plus, les règles sur les contenus violents ou explicites interdisent les contenus violents ou sanglants destinés à choquer les spectateurs ou à leur

	inspirer du dégoût, ainsi que ceux incitant à commettre des actes de violence. En particulier, YouTube interdit les types de contenus suivants : • Contenu incitant à commettre des actes de violence contre des individus ou un groupe de personnes en particulier • Vidéos, sons ou images incluant des accidents de la route, des catastrophes naturelles, des conséquences de guerre, des conséquences d'attaques terroristes, des combats de rue, des agressions physiques, des agressions sexuelles, des immolations, la torture, des cadavres, des manifestations ou émeutes, des vols, des procédures médicales ou des scènes similaires dont le but est de choquer les spectateurs ou de leur inspirer du dégoût Les règles de YouTube concernant l'incitation à la haine interdisent quant à elle tout contenu incitant à la violence ou à la haine contre des individus ou des groupes d'individus en fonction de l'une des caractéristiques suivantes : l'âge, la caste, le handicap, l'origine ethnique, l'identité et l'expression de genre, la nationalité, la race, le statut d'immigration, la religion, le sexe/genre, l'orientation sexuelle, le statut de victime d'un événement violent majeur ou de proche d'une victime, le statut d'ancien combattant. Les contenus qui encouragent la violence contre des individus ou des groupes d'individus en fonction de l'une des caractéristiques listées ci-dessus ou qui incitent à la haine contre des individus ou des groupes d'individus en fonction de l'une des caractéristiques listées ci-dessus sont interdits. Parmi les exemples proposés de contenus entrant dans le champ de cette catégorie, YouTube cite l'éloge ou l'apologie de la violence contre des individus ou des groupes en fonction des caractéristiques précitées. En juin 2019, YouTube a mis à jour ses règles concernant l'incitation à la haine de façon à interdire spécifiquement les vidéos alléguant qu'un groupe est supérieur dans le but de justifier des actes de discrimination, de ségrégation ou d'exclusion en fonction de caractéristiques telles que l'âge, le sexe/genre, l'origine ethnique, la religion, l'orientation sexuelle ou le statut de vétéran. YouTube a également annoncé qu'il retirerait les contenus niant l'existence attestée d'événements violents (Google, Youtube, 2019). Enfin, les règles concernant les contenus dangereux ou nuisibles interdisent les instructions pour tuer ou blesser, cela impliquant les contenus montrant aux spectateurs comment procéder pour tuer ou mutiler d'autres personnes, par exemple en fournissant des instructions pour fabriquer une bombe destinée à blesser ou à tuer des êtres humains. Elles interdisent également les contenus représentant des événements violents s'ils sont destinés à faire la promotion ou l'apologie de tragédies violentes, telles qu'une fusillade dans une école.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont	Le règlement de la communauté de YouTube est disponible sur https://www.youtube.com/about/policies/#community-guidelines. Les règles concernant les organisations criminelles violentes sont disponibles sur https://support.google.com/youtube/answer/9229472?hl=en&ref_topic=9282436. Les règles concernant les contenus violents ou explicites sont disponibles sur https://support.google.com/youtube/answer/2802008?hl=en-GB&ref_topic=9282436.

communiquées	Les règles concernant l'incitation à la haine sont disponibles sur https://support.google.com/youtube/answer/2801939?hl=en. Le règlement relatif aux contenus dangereux ou nuisibles est disponible sur https://support.google.com/youtube/answer/2801964?hl=en&ref_topic=9282436.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non. Le règlement de la communauté de YouTube s'applique aux vidéos, descriptions de vidéos, commentaires et diffusions en direct, ainsi qu'à tout autre produit ou toute autre fonctionnalité YouTube.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Si un convenu enfreint l'une des règles de YouTube concernant les contenus, YouTube le retire.
4.1 Notifications des	L'utilisateur est informé du retrait d'un contenu par courrier électronique, par notification sur son ordinateur de bureau ou son appareil mobile, ainsi que par l'émission d'une alerte dans les réglages de sa chaîne (Google/

suppressions ou des autres décisions de sanction	YouTube, 2020). Si le retrait du contenu se traduit par un « avertissement » (voir la section 6 ci-dessous), YouTube informe l'utilisateur : • du contenu qui a été retiré ; • des règles avec lesquelles le contenu était en infraction ; • des conséquences de l'avertissement sur sa chaîne ; • de ce qu'il peut faire ensuite.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Si un utilisateur reçoit un avertissement et estime que YouTube a commis une erreur, il peut introduire un recours (Google, Youtube, 2020). YouTube informe l'utilisateur de l'issue du recours par courrier électronique. Le recours peut avoir une des issues suivantes : • Si YouTube juge que le contenu respectait le règlement de la communauté, il le rétablit et retire l'avertissement de la chaîne de l'utilisateur. Si l'utilisateur introduit un recours contre un avertissement (voir la section 6 ci-dessous) et que le recours est reçu favorablement, l'infraction suivante se traduit par un avertissement. • Si YouTube juge que le contenu respectait le règlement de la communauté, mais qu'il ne convient pas à tous les publics, il applique une restriction d'âge. Si le contenu est une vidéo, elle ne sera pas accessible aux utilisateurs non connectés à leur compte, aux utilisateurs de moins de 18 ans ou à ceux dont le mode restreint (Google, Youtube, 2020) est activé. Si le contenu est une miniature personnalisée, il est retiré. • Si YouTube juge que le contenu enfreignait le règlement de la communauté, l'avertissement est maintenu et la vidéo reste inaccessible sur la plateforme. Les recours rejetés ne donnent pas lieu à l'imposition d'une sanction supplémentaire. Les utilisateurs peuvent introduire un seul et unique recours pour chaque avertissement.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de	YouTube fournit à ses utilisateurs des outils leur permettant de signaler les contenus qui enfreignent son règlement de la communauté (Google, Youtube, 2020). YouTube a mis au point des systèmes automatiques facilitant la détection de contenus susceptibles d'enfreindre ses règles. Lorsque ces systèmes repèrent un contenu potentiellement problématique, des évaluateurs vérifient s'il enfreint effectivement les règles de la plateforme. Dans l'affirmative, le contenu est supprimé et utilisé pour entraîner les systèmes automatisés et améliorer leurs résultats au fur et à mesure. Concernant plus particulièrement les systèmes automatisés qui détectent les contenus extrémistes (un terme qui n'est pas défini), le personnel de YouTube a examiné manuellement plus de deux millions de vidéos pour fournir des exemples aux systèmes. YouTube investit par ailleurs dans un réseau rassemblant plus de 180 universitaires, partenaires publics et ONG qui apportent leur expertise aux systèmes d'application des règles de la plateforme, en particulier dans le cadre du programme YouTube Trusted Flagger. (Google, Youtube, 2020)[46] Pour l'extrémisme violent, les membres de ce réseau comptent l'International Centre for the Study of Radicalisation du King's College de Londres (The International Centre for the Study of Radicalisation (ICSR), 2020), l'Institute for Strategic Dialogue (ISD Global,

partage d'empreintes numériques ou d'adresses URL)	n.d.), le Wahid Institute en Indonésie et des organismes publics spécialisés dans le contre-terrorisme. Les participants au programme Trusted Flagger reçoivent une formation sur la manière d'appliquer le règlement de la communauté, et, étant donné que leurs signalements présentent un taux de fiabilité plus élevé que ceux des utilisateurs classiques, ils sont examinés en priorité par les équipes de YouTube. Les contenus signalés par les participants au programme sont soumis aux mêmes règles que les signalements envoyés par d'autres utilisateurs et sont examinés par des équipes à même de décider s'ils enfreignent effectivement le règlement de la communauté. Les utilisateurs individuels, les organismes publics et les ONG peuvent participer au programme YouTube Trusted Flagger. Les participants doivent s'engager à régulièrement signaler des contenus susceptibles d'enfreindre le règlement de la communauté et à rester en permanence en contact avec YouTube pour faire remonter différentes informations en relation avec les types de contenus de la plateforme. YouTube indique que les discours haineux constituent un domaine où il est difficile d'appliquer des règles à grande échelle, les décisions nécessitant une compréhension nuancée des langues et des contextes locaux. Pour pouvoir mettre en œuvre de manière cohérente ses règles en la matière, YouTube a renforcé les compétences linguistiques et de ce domaine de son équipe d'évaluateurs. YouTube déploie également une technologie d'apprentissage automatique pour mieux détecter les contenus potentiellement haineux à soumettre à un contrôle humain, mettant à profit les leçons tirées de son expérience dans l'application des règles concernant d'autres types de contenus, tels que l'extrémisme violent (Google, Youtube, n.d.). Le coût économique marginal d'une utilisation des outils automatisés pour détecter les contenus terroristes et extrémistes violents est probablement très faible (bien que les coûts fixes puissent être élevés), alors que celui d'un recours à des modérateurs humains pour la même fonction est probablement relativement élevé. YouTube est un membre fondateur du GIFCT et participe au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	La première fois qu'un utilisateur publie un contenu qui enfreint les règles de la communauté, il reçoit une mise en garde et sa chaîne n'est pas sanctionnée. En cas de nouvelle infraction, YouTube envoie un avertissement à l'utilisateur. La chaîne est supprimée si l'utilisateur reçoit trois avertissements en 90 jours. Lors de l'émission du premier avertissement, l'utilisateur ne peut plus effectuer les opérations suivantes pendant une semaine : • Mettre en ligne des vidéos, des diffusions en direct ou des stories • Créer des miniatures personnalisées ou des publications destinées à la communauté • Créer, modifier ou ajouter des collaborateurs à des listes de lecture • Ajouter ou supprimer des listes de lecture à partir de la page de lecture d'une vidéo avec le bouton « Enregistrer »

	Tous les droits sont automatiquement restitués après une semaine, mais l'avertissement reste associé à la chaîne de l'utilisateur pendant 90 jours. Si l'utilisateur reçoit un deuxième avertissement dans les 90 jours suivant le premier avertissement, il ne pourra plus publier de contenu pendant deux semaines. S'il ne reçoit aucun nouvel avertissement, ses droits sont automatiquement rétablis à l'issue de ces deux semaines, mais chaque avertissement reste associé à sa chaîne pendant 90 jours. Si l'utilisateur reçoit trois avertissements au cours d'une seule et même période de 90 jours, sa chaîne est définitivement supprimée de YouTube (Google, YouTube, n.d.). Outre ce mécanisme reposant sur trois avertissements, une chaîne peut être clôturée en cas d'abus grave unique (comportement prédateur, par exemple) ou si elle a été créée spécialement pour enfreindre les règles de YouTube (c'est le cas des comptes destinés à envoyer des messages indésirables, par exemple). Lorsqu'une chaîne est clôturée, toutes ses vidéos sont supprimées. Lorsque des contenus n'enfreignent pas les règles de YouTube, mais s'approchent des critères de suppression et pourraient choquer certains spectateurs, certaines fonctionnalités peuvent être désactivées. Le contenu reste disponible sur YouTube, mais la page de lecture n'affiche plus de commentaires, de suggestions de vidéos ou de mentions « J'aime », et un message d'avertissement s'affiche avant la vidéo. Ces vidéos ne peuvent en outre pas être monétisées. La désactivation de fonctionnalités n'implique pas l'envoi d'un avertissement à la chaîne (Google, YouTube, n.d.). YouTube informe l'utilisateur de sa décision de désactiver certaines fonctionnalités par courriel. L'utilisateur peut faire appel de cette décision.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Oui (Google, n.d.). YouTube publie des rapports de transparence sur l'application de son règlement de la communauté. L'une des sections de ces rapports est consacrée à l'extrémisme violent (Google, YouTube, n.d.). Le dernier rapport de transparence précise que sont considérés comme ne respectant pas les règles de YouTube relatives à l'extrémisme violent les contenus produits par des organisations terroristes étrangères répertoriées comme telles par des gouvernements, sans indiquer cependant pas à quel(s) gouvernement(s) particulier(s) YouTube fait référence. Il indique aussi que YouTube interdit les contenus promouvant le terrorisme, tels que les contenus faisant l'apologie d'actes terroristes ou incitant à la violence. Il mentionne en outre que les contenus produits par des groupes extrémistes violents qui ne sont pas répertoriés comme organisations terroristes étrangères par des autorités sont souvent couverts par les règles de YouTube sur la publication de contenus haineux, violents ou choquants (voir la section 1 ci-dessus), y compris les contenus principalement destinés à choquer, ou à caractère sensationnel ou gratuit.
8. Informations ou types de données	YouTube publie : • le nombre de demandes de suppression de contenus émises par les autorités, réparties en six catégories (sécurité nationale,

figurant dans les rapports de transparence	diffamation, biens et services réglementés, confidentialité et sécurité, droits d'auteur, tous les autres) (Google, 2010-2020) ; • le nombre de chaînes supprimées, en fonction du motif de clôture (dont l'incitation à la violence et l'extrémisme violent) ; • le nombre de vidéos supprimées par source de détection initiale (détection automatique, signaleurs de confiance individuels, utilisateurs, organisations non gouvernementales et organismes publics) ; • le pourcentage de vidéos détectées initialement par les systèmes automatiques, avec ou sans aucune vue, c'est-à-dire le pourcentage de vidéos qui ont été supprimées avant d'avoir été vues par rapport à celles qui ont été supprimées après avoir obtenu quelques vues ; • le nombre et le pourcentage de signalements humains, par motif de signalement (dont l'incitation au terrorisme). YouTube indique qu'une vidéo peut être signalée plusieurs fois et pour différentes raisons, et que le signalement n'entraîne pas forcément un retrait. Les vidéos signalées par des personnes sont supprimées lorsqu'un modérateur qualifié confirme qu'elles ne respectent pas les règles (Google, Youtube, 2017-2020) ; • le nombre total de recours reçus que YouTube reçoit par trimestre pour des vidéos supprimées à la suite d'une violation des règles communautaires ainsi que le nombre total de vidéos que YouTube a réactivées par trimestre par suite d'un recours introduit après une suppression motivée par une violation des règles de la communauté. • le pourcentage et le nombre de vidéos supprimées selon le motif de suppression (dont le non-respect des règles de YouTube relatives à l'extrémisme violent et des règles concernant l'incitation à la haine) (Google, YouTube, n.d.) ; • le nombre de commentaires supprimés selon le motif de suppression (dont le non-respect des règles de YouTube relatives à l'extrémisme violent et des règles concernant l'incitation à la haine) ; • le pourcentage de commentaires supprimés en fonction de l'origine du premier signalement (détection automatique ou manuelle). Le rapport de transparence de YouTube comprend une section « Featured Policies » présentant le nombre total de vidéos supprimées pour non-respect de ses règles relatives à l'extrémisme violent et à l'incitation à la haine.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports	Aucune information n'est communiquée.

de transparence	
10. Fréquence de publication des rapports de transparence	La fréquence de publication est trimestrielle. Le dernier rapport de transparence porte sur le deuxième trimestre de 2020.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. Voir les sections 7 et 8 ci-dessus.

3. WhatsApp

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Les conditions d'utilisation de WhatsApp ne définissent pas les contenus terroristes et extrémistes violents. Cependant, la section « Sécurité et intégrité » des conditions d'utilisation indique que WhatsApp s'efforce de garantir la sécurité et l'intégrité de ses services en traitant de manière appropriée les personnes faisant preuve d'un comportement abusif et menant des activités contraires à ses conditions d'utilisation. Les termes « comportement abusif » et « activités contraires à ses conditions d'utilisation » peuvent s'appliquer aux utilisateurs diffusant des contenus terroristes et extrémistes violents, bien que ce ne soit pas précisé explicitement. Ces termes ne sont pas définis. Les conditions d'utilisation interdisent toute utilisation des services à mauvais escient, « tout comportement nuisible envers autrui » et toute violation des conditions et politiques. WhatsApp indique que les utilisateurs doivent consulter et utiliser ses services uniquement à des « fins légales, autorisées et acceptables », notamment ne pas les utiliser d'une manière qui soit de « nature illégale, obscène, diffamatoire, menaçante, intimidante, haineuse, racialement ou ethniquement offensante, assimilée à du harcèlement ou incite ou encourage un comportement illégal ou déplacé pour d'autres raisons, y compris la promotion de crimes violents ».
2. Manière dont les conditions d'utilisation ou les règles de la	Les textes sont disponibles sur https://www.whatsapp.com/legal/#terms-of-service.

communauté sont communiquées	
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non. WhatsApp ne propose pas de fonctionnalité de diffusion en direct qu'un invité puisse rejoindre.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	WhatsApp indique de façon générale qu'il peut modifier, suspendre ou résilier l'utilisation ou l'accès à ses services à tout moment en cas de comportement suspect ou illégal ou s'il estime raisonnablement que l'utilisateur ne respecte pas ses conditions d'utilisation ou crée un préjudice ou un risque de préjudice pour les utilisateurs ou d'autres personnes. Aucune procédure de recours n'est indiquée, mais si un utilisateur pense que son compte a été résilié ou suspendu par erreur, il peut contacter WhatsApp par courriel à l'adresse support@whatsapp.com.
4.1 Notifications des suppressions ou des autres décisions de sanction	Si un numéro est exclu, l'utilisateur en est informé conformément aux explications données à l'adresse https://faq.whatsapp.com/general/account-and-profile/seeing-the-message-your-phone-number-is-banned-from-using-whatsapp-contact-support-for-help/?lang=en.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Aucune procédure de recours n'est indiquée, mais si un utilisateur pense que son compte a été résilié ou suspendu par erreur, il peut contacter WhatsApp par courriel à l'adresse support@whatsapp.com.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	WhatsApp indique qu'il utilise la technologie avancée d'apprentissage automatique pour analyser les informations des groupes tels que les noms, photos de profil et descriptions de groupes afin d'améliorer sa capacité à détecter et supprimer les comportements abusifs et les activités susceptibles de porter préjudice à sa communauté et à la sécurité et à l'intégrité de ses services. Les signalements sont ensuite examinés par les modérateurs de WhatsApp, qui prennent des mesures appropriées, si nécessaire. WhatsApp indique aussi qu'il empêche certaines représentations au sein de discussions de groupes, comme l'utilisation de noms de groupes particuliers, afin de répondre aux exigences définies par la législation des États-Unis concernant les organisations désignées comme étant terroristes.

	Le coût économique marginal d'une utilisation des outils automatisés pour détecter les contenus terroristes et extrémistes violents est probablement très faible (bien que les coûts fixes puissent être élevés), alors que celui d'un recours à des modérateurs humains pour la même fonction est probablement relativement élevé. WhatsApp est membre du GIFCT.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	Si un utilisateur ne respecte pas les conditions d'utilisation et les règles de WhatsApp, ce dernier peut prendre des mesures à l'encontre du compte de l'utilisateur, comme le désactiver ou le suspendre. Dans ce cas, l'utilisateur ne doit pas créer un autre compte sans y être autorisé par WhatsApp. Si WhatsApp procède à la fermeture d'un groupe, ses participants ne seront plus en mesure d'envoyer de messages dans ce groupe. En outre, WhatsApp indique qu'il peut interdire aux administrateurs de ces groupes d'utiliser WhatsApp. WhatsApp indique encore que s'il a connaissance d'un comportement abusif ou d'activités contraires à ses conditions d'utilisation, il prendra des mesures appropriées en supprimant le compte ou les activités de l'utilisateur concerné ou en contactant les services chargés de l'application des lois.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Pas encore, mais la publication de rapports de transparence étant une condition de la participation au GIFCT, l'on peut s'attendre à ce que WhatsApp le fasse dans un avenir proche.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. Par exemple, après les attentats de Christchurch, deux extrémistes violents d'extrême droite qui faisaient apparemment partie d'un groupe WhatsApp intitulé « Christian White Militia » ont publié des déclarations encourageant le terrorisme en mars 2019 (Dearden, 2019).

4. Facebook Messenger

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition particulière des contenus terroristes et extrémistes violents. Facebook est toutefois l'un des rares services internet à disposer d'une définition bien précise du terrorisme et des termes associés dans ses « standards de la communauté ». Conformément à ce qui est indiqué au point 3 de ses conditions d'utilisation et au point 1.3 de ses politiques développeur, les standards de la communauté Facebook s'appliquent aussi à Messenger pour ce qui concerne le contenu généré par les utilisateurs ou les robots de messagerie. Dans la section des standards de la communauté Facebook intitulée « Individus et organisations dangereux » (Facebook, n.d.[1]), Facebook indique que les organisations ou individus qui revendiquent des objectifs violents ou qui sont impliqués dans des activités violentes ne sont pas les bienvenus sur Facebook. Ces organisations et individus sont définis comme celles et ceux qui sont impliqués dans les activités suivantes : • les activités terroristes ; • la haine organisée ; • les tueries (y compris les tentatives) ou homicides multiples ; • la traite des personnes ; • la violence ou les activités criminelles organisées. Les contenus soutenant ou faisant l'apologie de groupes, dirigeants ou individus impliqués dans ces activités font l'objet de la prise de mesures. Les individus (vivants ou décédés) et les groupes suivants ne peuvent pas maintenir une présence (par exemple, en possédant un compte, une page ou un groupe) sur la plateforme : organisations terroristes, terroristes, organisations animées par la haine (ainsi que leurs dirigeants et leurs membres prééminents) et les meurtriers de masse et de plusieurs homicides. Il n'est pas possible à quelqu'un d'avoir un compte Messenger sans compte Facebook. Les organisations terroristes et les terroristes comprennent les acteurs non étatiques qui : • prennent part à des actes de violence intentionnels et prémédités, les défendent ou les soutiennent de manière active ; • causent ou tentent de causer la mort, des blessures ou des dommages graves à des civils ou à toute autre personne ne prenant pas directement part aux hostilités dans le cadre d'un conflit armé, ou des dommages matériels graves associés à la mort, à des blessures graves ou à des dommages graves dont les victimes sont des civils ;

	• dans le but d'assujettir, d'intimider ou d'influencer une population, un gouvernement ou une organisation internationale ;
	• pour atteindre un objectif politique, religieux ou idéologique.
	
	Les organisations incitant à la haine sont définies comme des associations de trois personnes ou plus qui sont organisées sous un nom, un signe ou un symbole et dont l'idéologie, les déclarations ou les actions physiques portent atteinte à des individus en fonction de caractéristiques, notamment l'origine ethnique, l'affiliation religieuse, la nationalité, le genre, le sexe, l'orientation sexuelle, une maladie grave ou le handicap.
	
	Un homicide est considéré comme une tuerie s'il fait au moins trois victimes en un incident. Est considéré comme meurtrier en série tout individu ayant commis au moins deux meurtres au cours de plusieurs incidents ou à plusieurs endroits.
	
	Messenger interdit dans les images ou photos de profil des groupes Messenger les symboles qui représentent l'une des organisations ou l'un des individus cités ci-dessus s'ils ne sont pas partagés avec un contexte condamnant le contenu ou en discutant de façon neutre.
	
	Messenger prend également des mesures contre les utilisateurs dont il a appris qu'ils partageaient des contenus :
	• prônant l'une des organisations ou l'un des individus cités ci-dessus ou tout acte commis par eux ;
	• soutenant ou représentant des événements qu'elle désigne comme des attaques terroristes, des crimes de haine ou des meurtres de masse ;
	• prônant l'une des organisations ou l'un des individus cités ci-dessus ou tout acte commis par eux.
	
	Conformément à la section intitulée « Violence et provocation » des standards de la communauté Facebook (Facebook, n.d.[2]), Messenger prend également des mesures contre les utilisateurs dont Messenger a appris qu'ils partagent des propos qui incitent à la violence grave ou la facilitent. En particulier, les utilisateurs ne sont pas autorisés à partager :
	
	• des menaces pouvant entraîner la mort (ou toute autre forme de violence très grave) d'une cible, les menaces correspondant à l'un des éléments suivants :
	
	  • déclaration de l'intention de commettre un acte de violence très grave,
	  • appel à commettre un acte de violence très grave, y compris les contenus n'indiquant pas de cible précise, mais un symbole représentant la cible ou l'image d'une arme pour représenter la violence,
	  • déclaration prônant un acte de violence très grave,

	• déclaration intentionnelle ou conditionnelle de commettre un acte de violence très grave ; • des contenus qui demandent ou proposent des services de location d'un tueur (par exemple, tueur à gages, mercenaire, assassin) ou préconisent de faire appel à un tueur à gages, un mercenaire ou un assassin contre une cible ; • des reconnaissances, des déclarations d'intention ou des incitations, des appels à action ou des déclarations intentionnelles ou conditionnelles à enlever une cible ; • des menaces pouvant causer des blessures graves (violence de gravité moyenne) à des personnes, des personnalités publiques mineures, des personnes ou des groupes vulnérables, les menaces correspondant à l'un des éléments suivants : • déclaration de l'intention de commettre un acte de violence, • déclaration faisant l'apologie de la violence, • appel à commettre un acte de violence de gravité moyenne, y compris les contenus n'indiquant pas de cible précise, mais un symbole représentant la cible, • déclaration intentionnelle ou conditionnelle de commettre un acte de violence, • contenu concernant d'autres cibles que des personnes, des personnalités publiques mineures, des personnes ou des groupes vulnérables, et tout élément indiqué ci-après considéré comme crédible : o déclaration de l'intention de commettre un acte de violence, o appel à commettre un acte de violence, o déclaration faisant l'apologie de la violence, o déclaration intentionnelle ou conditionnelle de commettre un acte de violence ; • des menaces pouvant causer un préjudice physique (ou toute autre forme de violence de faible gravité) à des personnes (autosignalement obligatoire) ou des personnalités publiques mineures, les menaces correspondant à l'un des éléments suivants : • déclaration d'intention, • appel à action, • déclaration intentionnelle ou conditionnelle de commettre un acte de violence de faible gravité ou incitant à le faire ; • des représentations de personnes ou de personnalités publiques mineures qui ont été manipulées pour contenir

	des menaces de violence sous forme de texte ou d'image (cible, flèche, pistolet pointé sur la tête, etc.) ;
	• des contenus créés dans le but exprès de désigner une personne comme un membre appartenant à un groupe à risque précis et reconnaissable ;
	• des instructions sur la façon de fabriquer ou d'utiliser des armes s'il existe des preuves de l'intention de blesser gravement ou de tuer des personnes, telles que :
	  • des messages énonçant clairement cette intention,
	  • des photos ou des vidéos montrant ou simulant le résultat (blessure grave ou mort) dans le cadre des instructions,
	  • sauf si les contenus cités ci-dessus sont partagés dans un but d'autodéfense récréative, à des fins d'entraînement militaire, pour des jeux vidéo commerciaux ou pour relater des événements d'actualité (publiés sur une page ou avec un logo de nouvelles ou d'actualités) ;
	• des instructions sur la façon de fabriquer ou d'utiliser des explosifs, sauf si le contexte indique clairement que le contenu est publié à des fins non violentes (par exemple, jeux vidéo commerciaux, but scientifique/pédagogique clair, feux d'artifice ou matériel de pêche) ;
	• des contenus comportant des déclarations d'intention, des appels au passage à l'acte ou faisant l'apologie de la violence de forte ou moyenne gravité en raison de la tenue d'élections, de l'inscription d'électeurs ou des résultats d'une élection ;
	• de fausses informations contribuant à provoquer de la violence imminente ou des préjudices physiques ;
	• des appels à action, des déclarations de l'intention d'apporter des armes dans un lieu donné, tel qu'un lieu de culte, ou des encouragements à le faire.

2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.facebook.com/communitystandards/ ainsi que, pour les développeurs, sur https://developers.facebook.com/devpolicy/. Les conditions d'utilisation sont disponibles sur https://www.facebook.com/terms.php.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.

4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Voir la section 4 du profil de Facebook.
4.1 Notifications des suppressions ou des autres décisions de sanction	Voir la section 4.1 du profil de Facebook.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Voir la section 4.2 du profil de Facebook.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Voir la section 5 du profil de Facebook.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	Voir la section 6 du profil de Facebook.
7. Publication par le service de rapports de transparence sur les	Voir la section 7 du profil de Facebook.

contenus terroristes et extrémistes violents	
8. Informations ou types de données figurant dans les rapports de transparence	Voir la section 8 du profil de Facebook.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Voir la section 9 du profil de Facebook.
10. Fréquence de publication des rapports de transparence	Voir la section 10 du profil de Facebook.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. Voir les sections 7 et 8 du profil de Facebook.

5. iMessage/FaceTime

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition particulière des contenus terroristes et extrémistes violents. Cependant, les conditions générales des services Apple Media (qui s'appliquent à iMessage et FaceTime) interdisent aux utilisateurs de publier des contenus répréhensibles, insultants, illicites, trompeurs ou dangereux, qu'il s'agisse de commentaires, d'images, de photos, de vidéos ou de podcasts (y compris dans les métadonnées et les illustrations associées).
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.apple.com/ca/legal/internet-services/itunes/ca/terms.html.
3. Présence de dispositions précises applicables aux contenus diffusés en direct	Non.

dans les conditions d'utilisation ou les règles de la communauté	
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Aucune procédure n'est indiquée. Apple indique de manière générale qu'il peut surveiller et décider de retirer ou de modifier les contenus publiés.
4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Aucune procédure de recours n'est indiquée.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Apple a mis en place un système permettant aux utilisateurs de signaler les contenus qui enfreignent ses règles de publication (*Submission Guidelines*, figurant dans ses conditions générales des services Apple Media). Les signalements sont ensuite vérifiés et traités par l'équipe d'Apple. Dans la mesure où iMessage et FaceTime sont cryptés, il est difficile de comprendre comment un algorithme ou un évaluateur travaillant pour Apple pourrait détecter un contenu problématique et, notamment, un contenu terroriste et extrémiste violent. Le coût économique marginal d'un recours à des modérateurs humains pour détecter les contenus problématiques est probablement relativement élevé. Apple n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions	Si Apple détecte une violation potentielle ou avérée de l'une des dispositions de ses conditions générales, il peut, sans en avertir l'utilisateur au préalable, désactiver son identifiant Apple, sa licence d'utilisation des logiciels Appel ou son accès à ses services, qui comprennent iMessage et FaceTime.

d'utilisation ou des règles de la communauté	
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non. Apple publie des rapports de transparence (Apple, n.d.) qui comportent une section sur les demandes de suppression de contenu émanant des autorités et de tiers privés signalant des violations de ses conditions générales ou de la législation locale, mais qui ne mentionnent pas précisément les contenus terroristes et extrémistes violents.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	C'est possible. Un manuel sur la sécurité publié par l'EIIL recommandait d'utiliser iMessage pour protéger l'identité des sympathisants du mouvement, (Zetter, 2015), mais rien ne prouve que ces derniers l'ont effectivement utilisé (Dilger, 2015). Par ailleurs, le FBI a récemment réussi à déverrouiller l'iPhone de l'auteur de l'attaque de Pensacola et a découvert que ce dernier avait été en contact avec Al-Qaida « avec des applications chiffrées de bout en bout ». Il n'est toutefois pas précisé si iMessage ou FaceTime ont été effectivement utilisés (Sky News, 2020).

6. WeChat

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. Le code du bon usage de WeChat interdit aux utilisateurs de soumettre, de mettre en ligne, d'envoyer ou d'afficher des contenus qui, selon l'avis raisonnable de WeChat ou effectivement : • enfreignent des lois ou des réglementations (ou peuvent entraîner une violation de lois ou de réglementations) ; • créent un risque de perte ou de préjudice pour des personnes ; • nuisent à des personnes (adultes ou mineures) ou les exploitent de quelque façon que ce soit, notamment par des

	messages d'intimidation, de harcèlement ou des menaces de violence ; • sont haineux, assimilés à du harcèlement, insultants, racialement ou ethniquement offensants, diffamatoires, humiliants (publiquement ou de quelque façon que ce soit), menaçants, grossiers ou répréhensibles de quelque façon que ce soit.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.wechat.com/en/service_terms.html et https://www.wechat.com/en/acceptable_use_policy.html. (Tencent, n.d.)
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	WeChat indique qu'il peut examiner (mais ne s'engage pas à le faire) les contenus (dont les contenus publiés par ses utilisateurs) ou les programmes ou services tiers proposés sur WeChat pour vérifier s'ils respectent ses règles ainsi que les lois et réglementations en vigueur, ou s'ils ne sont pas répréhensibles de quelque manière que ce soit, et se réserve le droit de bloquer ou de supprimer des contenus pour quelque raison que ce soit, conformément aux lois et réglementations en vigueur.
4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.

4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Aucune procédure de recours n'est indiquée.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	WeChat ne fournit pas d'informations à cet égard. Les entreprises chinoises en ligne, dont WeChat, possèderaient une équipe de modérateurs chargée de traiter les contenus problématiques. [47] Les activistes politiques rapportent avoir été suivis à la suite de ce qu'ils ont dit sur WeChat et que des enregistrements de dialogues instantanés ont déjà été produits à titre de preuve au tribunal (Zhong, 2018). Par ailleurs, des études ont montré que WeChat fait appel à des algorithmes (Knockel J. L.-N., 2018), applique des mécanismes de filtrage par mots clés et recourt au blocage d'adresses universelles (Ruan L. J.-N., 2016) pour censurer les contenus qui enfreignent ses conditions d'utilisation (celles-ci pouvant également porter sur la publication de contenus terroristes et extrémistes violents). S'il a été rapporté que ces méthodes ne s'appliquaient qu'à des numéros de téléphone de Chine continentale, (Ruan L. J.-N., 2016), des études récentes ont révélé que les comptes internationaux (c'est-à-dire non chinois) faisaient également l'objet d'une surveillance « pour entraîner et construire le système de censure politique chinois de WeChat de manière invisible ». (Knockel, et al., 2020) Le coût économique marginal d'une utilisation des outils automatisés pour détecter les contenus problématiques est probablement très faible (bien que les coûts fixes puissent être élevés), alors que celui d'un recours à des modérateurs humains pour la même fonction est probablement relativement élevé. WeChat n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	WeChat indique qu'il peut suspendre ou résilier l'accès à ses services s'il pense raisonnablement qu'un utilisateur a enfreint ses conditions d'utilisation, que son utilisation du service crée un risque pour WeChat ou d'autres utilisateurs, que la suspension ou la résiliation est requise par la législation en vigueur ou à sa seule et absolue discrétion.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données	Sans objet.

figurant dans les rapports de transparence	
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. La fusillade de Christchurch a été publiée sur WeChat (Kenny, 2019). WeChat a également été utilisé pour diffuser de la propagande anti-musulmans (Huang, 2018).

7. Instagram

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Les règles de la communauté d'Instagram précisent qu'Instagram n'est pas un espace où soutenir ou faire l'éloge du terrorisme, du crime organisé ou de groupes haineux, ni où encourager la violence ou attaquer quiconque en raison de sa couleur de peau, de son origine ethnique, de sa nationalité, de son sexe, de son genre, de son identité de genre, de son orientation sexuelle, de son affiliation religieuse, de handicaps ou d'états pathologiques. Les menaces sérieuses d'atteintes à la sécurité publique et personnelle sont interdites, de même que le partage d'images visant à glorifier la violence. Instagram applique des règles de mise en œuvre communes, quoique non identiques, pour interpréter les règles de la communauté d'Instagram et les standards de la communauté Facebook. Pour obtenir de plus amples renseignements sur les règles de la communauté d'Instagram, les utilisateurs peuvent consulter les standards de la communauté Facebook, vers lesquels Instagram renvoie en lien en plusieurs endroits. Par exemple, les règles de la communauté d'Instagram précisent qu'« Instagram n'est pas un espace où soutenir ou faire l'éloge du terrorisme, du crime organisé ou de groupes haineux », et fournit un lien direct vers les standards de la communauté Facebook pour plus de détails sur les principes sous-tendant ces règles.
2. Manière dont les conditions	Les règles de la communauté d'Instagram sont disponibles sur https://help.instagram.com/477434105621119?helpref=page_content.

d'utilisation ou les règles de la communauté sont communiquées	Les conditions d'utilisation d'Instagram sont disponibles sur https://help.instagram.com/581066165581870.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Instagram peut supprimer un contenu s'il enfreint les règles de la communauté, ou désactiver ou résilier un compte.
4.1 Notifications des suppressions ou des autres décisions de sanction	Instagram avertit l'utilisateur concerné de la suppression d'un contenu ou de la suspension ou de la résiliation de son compte.
4.2 Mécanismes de recours en cas de suppression ou	Si l'utilisateur pense que son contenu a été supprimé ou que son compte a été résilié par erreur, il peut faire appel de cette décision. Il est possible de faire appel de la décision de supprimer un contenu considéré comme contraire aux politiques de « lutte contre le terrorisme » d'Instagram (qui ne sont pas précisées). Si le contenu a effectivement été supprimé par

d'autres décisions de sanction	erreur, Instagram le remettra en ligne et retirera l'infraction des archives du compte. En février 2020, Instagram a mis en place un mécanisme de recours simplifié pour les comptes désactivés, accessible directement depuis l'application, sans passer par son centre d'aide. Voir https://about.instagram.com/blog/announcements/safer-internet-day-2020/.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Instagram dispose d'une option de signalement intégrée qui permet aux utilisateurs de signaler les contenus qui enfreignent les règles de la communauté. Il possède aussi une équipe mondiale qui examine les signalements et supprime les contenus qui transgressent ses règles. Instagram déclare qu'il peut collaborer avec les services chargés de l'application des lois, en particulier s'il estime qu'un contenu représente un réel risque de préjudice physique ou une atteinte directe à la sécurité publique. Le document « Understanding the Community Standards Report » (Facebook) explique qu'Instagram applique les mêmes méthodes que Facebook pour identifier et retirer les contenus répréhensibles, y compris les contenus terroristes et extrémistes violents. Le coût économique marginal d'une utilisation des outils automatisés pour détecter les contenus terroristes et extrémistes violents est probablement très faible (bien que les coûts fixes puissent être élevés), alors que celui d'un recours à des modérateurs humains pour la même fonction est probablement relativement élevé. Instagram est membre du GIFCT et participe au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	Instagram peut supprimer des contenus ou des informations partagés par les utilisateurs s'il estime qu'ils enfreignent ses conditions d'utilisation et autres politiques (notamment ses règles de la communauté). Il peut aussi refuser ou arrêter immédiatement de fournir tout ou partie de ses services à un utilisateur (en particulier en résiliant ou en désactivant son compte) si celui-ci enfreint de manière claire, sérieuse ou répétée ses conditions d'utilisation et autres politiques (notamment ses règles de la communauté). Instagram a annoncé récemment une mise à jour de sa politique relative à la désactivation des comptes. Outre la suppression des comptes comportant un certain pourcentage de contenus en infraction (non communiqué), il supprimera aussi les comptes qui ont enregistré un certain nombre d'infractions pendant une période donnée (non communiqué non plus) (Instagram, 2019).
7. Publication par le service de rapports de transparence sur les contenus	Oui. Le dernier rapport d'application des règles de la communauté de Facebook (Community Standards Enforcement Report, daté du deuxième trimestre de 2020) comprend des informations d'Instagram sur les thèmes suivants : nudité des adultes et activité sexuelle, intimidation et harcèlement, nudité des enfants et exploitation sexuelle, biens réglementés, suicide et blessures auto-infligées, contenus violents et

terroristes et extrémistes violents	explicites, propos haineux, et organisations dangereuses : terrorisme et haine organisée.
8. Informations ou types de données figurant dans les rapports de transparence	Le thème « organisations dangereuses : terrorisme et haine organisée » regroupe trois types d'informations : - *Incidence (incidence des violations liées aux contenus terroristes et aux contenus violents et explicites sur Instagram)* L'indicateur d'incidence correspond au pourcentage de vues associées à des contenus en infraction du fait du lieu au terrorisme. Ainsi, la limite supérieure estimée par Instagram est de 0,05 % de vues de contenus enfreignant ses règles en matière de terrorisme au deuxième trimestre de 2020. En d'autres termes, sur 10 000 vues entrant dans le champ d'application de la politique d'Instagram relative au terrorisme, au maximum 5 présentaient des contenus en infraction. (Ces chiffres sont ceux après détermination définitive du statut des contenus, sans tenir compte des contenus initialement signalés comme enfreignant potentiellement la règle, mais jugés admissibles ultérieurement.) - *Contenus pour lesquels des mesures ont été prises (quantité de contenus sur lesquels Instagram est intervenu)* Les mesures prises peuvent comprendre la suppression d'un contenu d'Instagram, l'ajout d'un filtre d'avertissement sur des photos ou des vidéos susceptibles de déranger certains publics ou la désactivation d'un compte. Les contenus pour lesquels des mesures ont été prises désignent le nombre total d'éléments de contenu pour lesquels Instagram a pris des mesures pendant une période de rapport donnée au titre qu'ils enfreignaient ses règles de la communauté (dans le cas présent, les règles relatives aux contenus relevant du terrorisme). - *Taux de proactivité (pourcentage de contenus en infraction pour lesquels des mesures ont été prises par Instagram avant que les utilisateurs les signalent)* Ce chiffre indique le pourcentage de contenus pour lesquels Instagram a pris des mesures parce qu'ils enfreignaient ses règles avant qu'ils soient signalés par les utilisateurs. Il comptabilise les détections opérées à la fois par les outils d'intelligence artificielle et par les évaluateurs humains de Facebook. - *Recours (nombre de recours contre des décisions prises par Instagram à l'encontre de contenus en infraction)* Cet indicateur indique le nombre total de contenus pour lesquels des mesures ont été prises et qui ont fait l'objet d'une demande de réexamen sur la période couverte par le rapport. - *Contenus rétablis (nombre de contenus qu'Instagram a rétablis après les avoir retirés)* Le nombre de contenus rétablis correspond au nombre d'éléments de contenu qu'Instagram a rétablis au cours de la période considérée après avoir pris des mesures à leur encontre.

	Instagram fait également état des tendances récentes concernant les contenus relevant de la haine organisée et du terrorisme pour lesquels des mesures ont été prises. Ainsi, son dernier rapport de transparence note que les contenus relevant de la haine organisée pour lesquels des mesures ont été prises ont augmenté 175 100 au premier trimestre de 2020 à 266 000 au deuxième trimestre de la même année, tandis que les contenus pour lesquels des mesures ont été prises au titre de la lutte contre le terrorisme ont diminué au deuxième trimestre de 440 600 à 388 800.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	- *Incidence.* Ce chiffre indique le nombre estimé de vues d'un contenu enfreignant les règles divisé par le nombre estimé de vues de la totalité des contenus sur Instagram sur la période couverte par le rapport. Par exemple, une incidence d'organisations dangereuses de 0,18 % à 0,20 % signifie que sur 10 000 vues de contenus, 18 à 20, en moyenne, correspondaient à des vues de contenus enfreignant les normes d'Instagram concernant les organisations dangereuses. L'incidence fournit une indication de la fréquence à laquelle est vu un contenu interdit, et non le nombre total de ce type de contenus. Pour les violations relevant du terrorisme, en particulier, Instagram fournit uniquement une estimation de la limite supérieure, parce qu'il se dit « convaincu que l'incidence des vues des contenus enfreignant ces règles est inférieure à cette limite » (Facebook). - *Contenus pour lesquels des mesures ont été prises.* Ce chiffre indique le nombre total de contenus pour lesquels Instagram a pris des mesures au cours de la période couverte par le rapport parce qu'ils ne respectaient pas ses règles. Instagram ne comptabilise pas les contenus qui ont été transmis à un service chargé de l'application des lois. Cet indicateur comprend tant les contenus pour lesquels Instagram a pris des mesures après signalement par un utilisateur que ceux qu'Instagram a repérés de manière proactive. - *Taux de proactivité.* Cet indicateur est le résultat de la division du nombre d'éléments de contenu pour lesquels des mesures ont été prises à la suite d'une détection par Instagram avant que les utilisateurs les signalent par le nombre total d'éléments de contenu pour lesquels des mesures ont été prises. - *Recours.* Cet indicateur indique le nombre total de contenus pour lesquels des mesures ont été prises et qui ont fait l'objet d'une demande de réexamen sur la période couverte par le rapport. Instagram fait observer que cet indicateur mesure le nombre d'éléments de contenu ayant fait l'objet d'un recours dans le trimestre, le nombre de contenus rétablis pendant le trimestre faisant quant à lui l'objet d'un autre indicateur (voir ci-dessous). Comme certains éléments de contenu peuvent avoir été rétablis au trimestre suivant et que certains contenus rétablis peuvent avoir fait l'objet d'un recours au trimestre précédent, une comparaison directe de ces indicateurs n'est pas possible. - *Contenus rétablis.* Pour calculer cet indicateur, Instagram compte le nombre d'éléments de contenu rétablis au cours de la

	période considérée après avoir pris des mesures à leur encontre. Instagram peut rétablir des contenus soit dans le cadre d'un recours introduit à la suite d'une décision de retrait, soit lorsqu'une raison de le faire parvient à la connaissance de Facebook.
10. Fréquence de publication des rapports de transparence	Les rapports de transparence d'Instagram sont publiés conjointement avec ceux de Facebook et selon le même calendrier.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. Les médias ont fait état de multiples exemples, (Carmen, 2015) (Hymas, 2019) (Cox, 2019).

8. TikTok

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition particulière des contenus terroristes et extrémistes violents. Les règles communautaires de TikTok précisent toutefois que les « personnes ou organisations dangereuses » ne peuvent pas utiliser la plateforme pour promouvoir le terrorisme, le crime ou d'autres types de comportements susceptibles de porter préjudice. Cette catégorie comprend expressément les terroristes et les organisations terroristes. TikTok définit les terroristes et organisations terroristes comme des acteurs non-étatiques qui utilisent la violence préméditée ou la menace de violence pour causer des dommages à des civils, pour intimider ou menacer une population, un gouvernement ou une organisation internationale pour atteindre des objectifs politiques, religieux, ethniques ou idéologiques. Plus globalement, TikTok définit les personnes ou organisations dangereuses comme les auteurs de crimes ou d'autres types de dommages graves. Ces groupes et crimes comprennent notamment les groupes haineux, les organisations extrémistes violentes, les homicides, la traite des personnes, le trafic d'organes, le trafic d'armes, le trafic de drogue, les enlèvements, les extorsions, le chantage, le blanchiment d'argent, la fraude, la cybercriminalité. Les noms, symboles, logos, slogans, uniformes, gestes, portraits ou autres objets destinés à représenter des

	personnes ou des organisations dangereuses, ou des contenus qui font l'éloge de personnes ou d'organisations dangereuses, les glorifient ou les soutiennent, sont interdits sur TikTok, à l'exception des contenus pédagogiques, historiques, satiriques, artistiques ou clairement identifiés comme des contre-discours ou qui visent à sensibiliser aux dommages causés par des personnes ou organisations dangereuses. En outre, TikTok interdit les « contenus violents et explicites », c'est-à-dire les contenus dont le caractère macabre ou choquant est poussé à l'extrême, en particulier ceux promouvant ou glorifiant des actes de violence abjects ou des souffrances provoquant le dégoût. Certaines exceptions sont admises, notamment les contenus publiés à titre d'actualités ou ceux visant à sensibiliser le public à certaines causes. Les contenus choquants sans motif légitime, à caractère sadique ou particulièrement explicites sont par exemple ceux représentant des morts violentes ou accidentelles impliquant des personnes réelles, des restes de corps humains démembrés, mutilés, carbonisés ou brûlés, des scènes sanglantes attirant l'attention sur des plaies ou des blessures ouvertes ainsi que des actes de violence physique graves. De même, TikTok interdit les contenus représentant des attaques ou des incitations à la violence contre des personnes ou des groupes de personnes en raison de caractéristiques faisant l'objet d'une protection particulière, y compris les propos haineux. Sont visés ici les contenus constituant des menaces verbales ou physiques d'actes de violence ou représentant des préjudices occasionnés à une personne ou à un groupe en raison d'une des caractéristiques protégées suivantes : couleur de peau, origine ethnique, nationalité, religion, caste, orientation sexuelle, sexe, genre, identité de genre, pathologie ou handicap grave et statut d'immigration. TikTok interdit encore les contenus déshumanisant ou incitant à la violence ou à la haine contre des personnes ou des groupes en raison des caractéristiques précitées, y compris les contenus alléguant d'une infériorité physique ou morale et appelant à la violence contre ces personnes ou ces groupes, ou justifiant une telle violence, les traitant de criminels, les comparant à des animaux, à des objets ou à d'autres entités non humaines, et promouvant ou justifiant leur exclusion, la ségrégation ou leur discrimination. Enfin, TikTok interdit les contenus intégrant des éléments idéologiques incitant à la haine (ceux-ci n'étant pas définis), y compris les contenus faisant la promotion d'idéologies incitant à la haine par des discours positifs ou la représentation de logos, de

	symboles, de drapeaux, de slogans, d'uniformes, de saluts, de gestes, de portraits, d'illustrations ou de noms de personnes associées à ces idéologies. Cette interdiction porte également sur les contenus niant que des événements violents attestés aient eu lieu ainsi que sur les musiques ou paroles faisant la promotion d'idéologies de la haine.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.tiktok.com/en/terms-of-use#terms-eea et https://www.tiktok.com/community-guidelines?lang=en.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	TikTok indique de manière générale qu'il peut à tout moment et sans notification préalable supprimer ou désactiver l'accès à un contenu à sa seule discrétion, pour quelque raison que ce soit ou sans raison. TikTok peut supprimer un contenu qu'il estime répréhensible, contraire à ses conditions d'utilisation ou à ses règles communautaires ou préjudiciable de quelque façon que ce soit à ses services ou à ses utilisateurs.
4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Si un utilisateur pense que son contenu a été supprimé par erreur, il peut faire appel de cette décision.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	TikTok combine plusieurs techniques et méthodes de modération de contenus pour identifier et retirer les contenus et les comptes qui enfreignent ses règles. Sur le plan technologique, TikTok a mis au point des systèmes signalant automatiquement certains types de contenus susceptibles d'enfreindre ses règles de la communauté. Ces systèmes sont entraînés pour identifier des types de comportements ou des signaux comportementaux et les signaler comme des contenus potentiellement interdits, de telle sorte que TikTok puisse prendre des mesures rapides et réduire le préjudice potentiel. TikTok fait observer qu'il suit attentivement les tendances, les développements du monde universitaire et les bonnes pratiques du secteur pour améliorer en permanence ses systèmes.

	En matière de modération de contenus, TikTok ne se fie pas uniquement à ses systèmes, qui restent perfectibles, pour vérifier le bon respect de ses règles. Ainsi, le contexte peut être primordial pour déterminer si certains contenus, notamment satiriques, sont en infraction. TikTok a ainsi mis en place une équipe de modérateurs qualifiés pour faciliter l'examen et le retrait des contenus ne respectant pas ses normes. Dans certains cas, l'équipe retire de manière proactive des contenus se rapprochant de l'infraction, comme les défis dangereux ou la désinformation préjudiciable. TikTok modère également les contenus à la suite de rapports reçus de ses utilisateurs. La fonctionnalité de signalement intégrée à l'application TikTok permet à l'utilisateur de choisir parmi différents motifs pour indiquer pourquoi il pense qu'un contenu peut enfreindre les règles de TikTok (comme un acte de violence, un préjudice, un harcèlement ou des propos haineux). Si les modérateurs de TikTok jugent qu'il y a infraction, le contenu est retiré. TikTok fait également appel à un panel d'experts de confiance pour l'aider à comprendre l'environnement dynamique des règles, et élabore des politiques ainsi que des stratégies de modération pour faire face aux contenus et aux comportements problématiques à mesure qu'ils se font jour. Ces experts comprennent notamment les huit experts du conseil consultatif américain de TikTok en matière de contenus (U.S. Content Advisory Council) ainsi que des organisations telles que ConnectSafely.org, le National Center for Missing and Exploited Children et WePROTECT Global Alliance (TikTok, 2019-2020). Le coût économique marginal d'une utilisation des outils automatisés pour détecter les contenus terroristes et extrémistes violents est probablement très faible (bien que les coûts fixes puissent être élevés), alors que celui d'un recours à des modérateurs humains pour la même fonction est probablement relativement élevé. TikTok n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	La violation des règles communautaires peut entraîner la suspension du compte, sa résiliation ou la suppression du contenu.
7. Publication par le service de rapports de transparence sur les	Non, pas à proprement parler. Toutefois, dans son deuxième rapport de transparence, TikTok a indiqué

contenus terroristes et extrémistes violents	qu'à la fin 2019, il a commencé à déployer une nouvelle infrastructure de modération de contenus permettant d'expliquer les raisons pour lesquelles des vidéos sont retirées de TikTok de manière plus transparente. Cette infrastructure permet à TikTok d'associer à une vidéo qui enfreint ses règles un libellé mentionnant la ou les règles qu'elle enfreint lorsqu'il la retire. Une seule et même vidéo peut donc être rattachée à plusieurs catégories de règles, y compris celle des organisations dangereuses, qui inclut les terroristes et les organisations terroristes (et, par extension, les contenus terroristes et extrémistes violents).
8. Informations ou types de données figurant dans les rapports de transparence	TikTok n'a fourni des indicateurs que pour le mois de décembre 2019, date à laquelle sa nouvelle infrastructure de modération de contenus est entrée en action. TikTok a fait état en particulier des indicateurs suivants : - Pourcentage de vidéos retirées pour non-respect des règles des catégories suivantes : o nudité des adultes et activité sexuelle ; o sécurité des mineurs ; o activités illégales et biens réglementés ; o suicide, blessures auto-infligées et actes dangereux ; o contenu violent et explicite ; o intimidation et harcèlement ; o propos haineux, respect de l'intégrité et de l'authenticité, et personnes et organisations dangereuses. - Nombre de vidéos retirées à l'échelle mondiale pour avoir enfreint les règles communautaires ou les conditions d'utilisation de TikTok - Pourcentage de vidéos repérées proactivement et retirées par les systèmes de TikTok avant qu'elle soient signalées par les utilisateurs - Pourcentage de vidéos retirées avant tout visionnage
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Aucune information n'est communiquée.
10. Fréquence de publication des rapports de transparence	Les rapports font l'objet d'une édition semestrielle.

11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui, voir les sections 7 et 8 ci-dessus.

9. QQ

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. Les conditions d'utilisation de QQ interdisent toutefois aux utilisateurs de soumettre, de mettre en ligne, d'envoyer ou d'afficher des contenus qui, selon l'avis raisonnable de QQ ou effectivement : • enfreignent des lois ou des réglementations (ou peuvent entraîner une violation de lois ou de réglementations) ; • créent un risque de perte ou de préjudice pour des personnes ; • nuisent à des personnes (adultes ou mineures) ou les exploitent de quelque façon que ce soit, notamment par des messages d'intimidation, de harcèlement ou des menaces de violence ; • sont haineux, assimilés à du harcèlement, insultants, racialement ou ethniquement offensants, diffamatoires, humiliants (publiquement ou de quelque façon que ce soit), menaçants, grossiers ou répréhensibles de quelque façon que ce soit.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.tencent.com/en-us/zc/termsofservice.shtml et https://www.tencent.com/en-us/zc/acceptableusepolicy.shtml[48].
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de	QQ indique qu'il peut examiner (mais ne s'engage pas à le faire) les contenus (dont les contenus postés par ses utilisateurs) ou les services tiers proposés sur QQ pour vérifier s'ils respectent ses règles ainsi que les lois et réglementations en vigueur, ou s'ils ne sont pas répréhensibles de quelque manière que ce soit, et se réserve le droit de bloquer ou de retirer des contenus pour quelque raison que ce soit, conformément aux lois et réglementations en vigueur.

procédures de recours contre ces décisions	
4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Aucune procédure de recours n'est indiquée.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Pas d'informations communiquées. QQ n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	QQ peut suspendre ou résilier l'accès à ses services s'il pense raisonnablement qu'un utilisateur a enfreint ses conditions d'utilisation, que l'utilisation de ses services crée un risque pour la plateforme ou d'autres utilisateurs, que la suspension ou la résiliation est requise par la législation en vigueur ou à sa seule et absolue discrétion.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.

10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Information inconnue.

10. Youku Tudou

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. Cependant, les conditions d'utilisation de Youku Tudou interdisent les contenus qui incitent à la haine et à la discrimination ethnique, ou qui menacent l'unité ethnique, ainsi que les contenus qui incitent à commettre des délits, glorifient la violence ou prônent des activités terroristes.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur http://mapp.youku.com/service/agreement-eng.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Youku Tudou indique « gérer » les informations mises en ligne, publiées ou envoyées par les utilisateurs et prendre des mesures, telles que suspendre l'envoi, supprimer des contenus mis en ligne pour empêcher leur diffusion, conserver des enregistrements et effectuer des signalements auprès des autorités compétentes si les informations mises en ligne sont interdites par les lois et réglementations en vigueur ou enfreignent les conditions d'utilisation.

4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Aucune procédure de recours n'est indiquée.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Youku Tudou ne fournit pas d'informations à cet égard. Youku Tudou n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	Le non-respect des conditions d'utilisation de Youku Tudou peut entraîner le retrait du contenu, le blocage du contenu et des informations, la suspension, la résiliation ou l'annulation du compte de l'utilisateur ou toute autre mesure conforme aux réglementations en vigueur.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.

11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Information inconnue.

11. Weibo

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition particulière des contenus terroristes et extrémistes violents. Les conditions d'utilisation de Weibo interdisent toutefois aux utilisateurs de mettre en ligne, d'afficher et d'envoyer des contenus offensants, insultants, intimidants, racialement discriminants, malveillants, violents ou illicites de quelque manière que ce soit.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.weibo.com/signup/v5/protocol.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Weibo indique de manière générale que ses opérateurs ont le droit d'examiner, surveiller et traiter le comportement et les informations des utilisateurs de la plateforme, notamment les informations sur les utilisateurs (informations sur le compte, données personnelles, etc.), les données de contenu (lieu, texte, photos, audio, vidéos, marques, brevets, publications, etc.) et le comportement des utilisateurs (relations, commentaires, lettres privées, sujets et activités auxquels ils participent, données de marketing, réclamations, etc.).
4.1 Notifications des suppressions ou des	Aucune notification n'est indiquée.

autres décisions de sanction	
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Aucune procédure de recours n'est indiquée.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Weibo a mis en place un système permettant aux utilisateurs de signaler les contenus illicites ou répréhensibles. Les signalements sont ensuite vérifiés et traités par des modérateurs. Le coût économique marginal d'un recours à des modérateurs humains pour détecter les contenus répréhensibles est probablement relativement élevé. Weibo n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	En cas de violation des conditions d'utilisation, Weibo est autorisé à suspendre ou résilier la fourniture de ses services.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.

11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. La fusillade de Christchurch a été publiée sur Weibo (Kenny, 2019).

12. QZone

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. Les conditions d'utilisation de QQ International[49] interdisent toutefois aux utilisateurs de publier, d'envoyer, de diffuser ou d'enregistrer des contenus contraires à la législation ou inappropriés, insultants, obscènes ou violents.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://imqq.com/html/FAQ_en/html/Miscellaneous_1.html. [50]
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Aucune procédure n'est précisée.
4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Aucune procédure de recours n'est indiquée.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les	QQ International ne fournit pas d'informations à cet égard.

utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	QQ International n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	QQ International indique qu'en cas de violation de ses conditions d'utilisation, il est autorisé à suspendre la licence de l'utilisateur, interrompre la fourniture de ses services, appliquer des restrictions de service, reprendre le compte de l'utilisateur, mener des enquêtes légales et d'autres mesures appropriées, en fonction de la gravité du comportement de l'utilisateur, et ce sans en avertir ce dernier au préalable.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Information inconnue.

13. iQIYI

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. Les conditions d'utilisation d'iQIYI interdisent toutefois la promotion du terrorisme, de l'extrémisme (sans mentionner spécifiquement l'extrémisme violent), de la haine, de la discrimination ethnique et la diffusion de la violence.
2. Manière dont les conditions d'utilisation ou	Les textes sont disponibles sur https://www.iqiyi.com/user/register/protocol.html.

les règles de la communauté sont communiquées	
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	iQIYI indique se réserver le droit d'annuler l'accès des utilisateurs à ses produits et services ou leur capacité à créer, mettre en ligne, publier et diffuser des contenus sans préavis.
4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Aucune procédure de recours n'est indiquée.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes	iQIYI ne fournit pas d'informations à cet égard. iQIYI n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.

numériques ou d'adresses URL)	
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	iQIYI précise qu'en cas de violation de ses conditions d'utilisation, il est autorisé à suspendre ou à annuler le compte concerné, et peut signaler le cas échéant les infractions aux autorités.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Information inconnue.

14. Reddit

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de	Il n'existe pas de définition particulière des contenus terroristes et extrémistes violents. La « politique de contenu » de Reddit interdit toutefois les contenus qui encouragent les actes de violence, en font l'apologie, incitent ou appellent à la violence ou à des préjudices physiques contre une personne ou un groupe de personnes.

conduite ou les règles de la communauté	De même, conformément à son rapport de transparence, Reddit retire les contenus terroristes.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.redditinc.com/policies/user-agreement et https://www.redditinc.com/policies/content-policy. Il convient de préciser que Reddit utilise un système de modération à plusieurs niveaux. La politique de contenu visée ci-dessus régit tous les contenus mis en ligne sur la plateforme Reddit, qui rassemble des milliers de communautés créées et modérées par les utilisateurs eux-mêmes, à titre volontaire. En plus de cette politique de contenu, qui s'applique à l'ensemble du site, les modérateurs définissent les règles de leurs communautés, en fonction des thèmes abordés par celles-ci. Ces règles sont clairement indiquées dans un encadré affiché sur la page de chaque communauté.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Oui. Les textes sont disponibles sur https://www.redditinc.com/policies/broadcasting-content-policy. Les contenus diffusés en direct sur Reddit sont soumis à des règles supplémentaires à celles figurant dans la politique de contenu. **Pas de contenu « inapproprié au travail » (NSFW)** Les diffusions sur Reddit ne doivent pas comprendre d'éléments « inappropriés au travail » (NSFW, *Not Safe for Work*). Comme l'indique la politique de contenu, cette mention désigne des contenus suggestifs ou comprenant de la nudité, de la pornographie ou de la violence, qu'un spectateur raisonnable ne souhaite pas nécessairement voir dans un endroit public ou formel tel que son lieu de travail. **Pas de comportement illégal ou dangereux** Les diffusions ne doivent pas représenter des activités illégales ou qui présentent un risque déraisonnable de dommage physique pour le sujet filmé ou les personnes à proximité. **Pas de contenu soumis à quarantaine** Les diffusions sur Reddit ne doivent pas comprendre d'éléments qui seraient sinon soumis à quarantaine. Comme l'indique la politique de contenu, ce sont des éléments que les utilisateurs moyens de la plateforme pourraient trouver très insultants ou choquants, ou qui encouragent les canulars.
4. Politiques et procédures de mise en œuvre et d'application des conditions	Sur le site, les administrateurs (des employés payés par Reddit) appliquent différentes méthodes pour faire appliquer les règles, telles que : • prier poliment l'utilisateur de retirer son contenu ;

d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	• prier moins poliment l'utilisateur de retirer son contenu ; • suspendre un compte de manière temporaire ou définitive ; • supprimer certaines options privilégiées d'un compte ou le restreindre ; • appliquer des restrictions à des communautés Reddit, comme ajouter la mention « NSFW » ou mettre en quarantaine (voir plus bas) ; • retirer des contenus ; • interdire des communautés Reddit. Outre les mesures de modération que les administrateurs Reddit peuvent prendre au niveau du site, les modérateurs bénévoles disposent d'un certain nombre de moyens pour appliquer les règles à l'échelle de leurs communautés. Ils peuvent par exemple exclure un utilisateur de la communauté (à titre temporaire ou définitif) ou en retirer ses publications. Ces mesures sont prises indépendamment des administrateurs Reddit. La mise en quarantaine (Reddit Inc., n.d.) est appliquée aux communautés (soit, *grosso modo*, des groupes partageant un intérêt commun) que les utilisateurs pourraient trouver insultantes ou choquantes ou qui visent à encourager les canulars, nécessitant par là une surveillance supplémentaire. L'objectif de cette mesure est de faire en sorte que le contenu de ces communautés ne soit pas vu accidentellement par des personnes qui ne le souhaiteraient pas consciemment ou sans un contexte approprié. Les communautés soumises à quarantaine affichent un message d'avertissement demandant explicitement à l'utilisateur de confirmer qu'il choisit de voir ce contenu. Elles ne génèrent aucune recette, n'apparaissent pas dans les fils d'actualité accessibles sans compte Reddit (par exemple les publications populaires) et ne figurent pas dans les résultats de recherches ou les recommandations. Reddit peut également appliquer d'autres restrictions déjà prévues ou susceptibles d'être développées à l'avenir (comme supprimer des outils de personnalisation).
4.1 Notifications des suppressions ou des autres décisions de sanction	Lorsque les administrateurs prennent des mesures d'application des règles, les utilisateurs en sont avertis. En cas de suspension de compte à l'échelle du site, une notification est envoyée par message privé. Un rappel visuel s'affiche également sur chaque page que l'utilisateur visite pendant la durée de la suspension ainsi que chaque fois qu'une action interdite est tentée, comme une publication ou le dépôt d'un commentaire. Lorsque des mesures sont prises à l'encontre d'un compte individuel, comme une suspension à l'échelle du site, l'utilisateur reçoit un message l'avertissant de la suspension et du motif de celle-ci. L'utilisateur suspendu verra également s'afficher directement à l'écran une banderole l'avertissant de la suspension.

	Si la suspension d'un utilisateur intervient à la suite de l'action d'un modérateur volontaire d'un *sous-reddit*, la notification s'effectue également par message privé. De plus amples informations sur la suspension des comptes sont disponibles à l'adresse https://www.reddithelp.com/hc/en-us/articles/360045734511-My-account-was-suspended-for-violating-Reddit-s-Content-Policy. Lorsque des éléments individuels de contenu sont retirés, ils sont remplacés par une « pierre tombale » pour indiquer aux visiteurs que des contenus précédemment disponibles ont été retirés. Lorsqu'un sous-reddit entier est retiré, une page de type pierre tombale informe les visiteurs du retrait effectué et de la règle enfreinte.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Les mesures prises par Reddit à l'encontre d'éléments individuels de contenu, d'un compte ou d'un sous-reddit entier à la suite d'une infraction à la politique de contenu peuvent faire l'objet d'un recours (l'« appel »), formé en remplissant un simple formulaire, disponible sur https://www.reddit.com/appeals. Les recours sont évalués par des employés Reddit et sont soit admis (avec rétablissement du contenu, du compte ou du sous-reddit) ou rejeté. Une procédure de recours spéciale s'applique aux sous-reddits placés en quarantaine. Pour faire sortir un sous-reddit de quarantaine, les modérateurs de la communauté (voir la section 5 ci-dessous) peuvent introduire un recours. Celui-ci devra comprendre un compte rendu détaillé des changements apportés aux pratiques de modération de la communauté (ces changements peuvent différer selon les communautés et peuvent rassembler diverses techniques, telles que l'ajout de modérateurs, la mise en place de nouvelles règles, l'utilisation d'outils de modération plus agressifs, la modification du type de la communauté, etc.). Il devra aussi apporter des preuves de l'exécution cohérente et durable de ces changements pendant au moins un mois, afin de montrer une évolution réelle de la communauté. Reddit peut à son entière discrétion supprimer ou retirer un contenu à tout moment et pour quelque raison que ce soit, notamment en cas de violation de ses conditions d'utilisation ou de sa politique de contenu, ou s'il doit autrement assumer la responsabilité du contenu. Il est possible de faire appel des mesures prises par Reddit en cas de violation de sa politique de contenu, qu'elles concernent un compte individuel ou une communauté. Ce sont les employés de Reddit qui examinent les appels.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par	Reddit s'appuie sur un système de modérateurs bénévoles parmi ses utilisateurs. La modération d'une communauté Reddit est une fonction non officielle et non rémunérée. Les créateurs d'une communauté en deviennent automatiquement les

exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	premiers modérateurs et ils peuvent nommer d'autres utilisateurs pour les aider. Reddit se réserve le droit de supprimer ou de limiter la capacité d'un utilisateur à exercer la fonction de modérateur à tout moment et pour quelque raison que ce soit, notamment en cas de violation de ses conditions d'utilisation. Les modérateurs doivent respecter la politique de modération (Reddit Inc., 2017), et, s'ils reçoivent un signalement concernant leur communauté, ils doivent prendre des mesures de modération en supprimant le contenu ou en le transmettant aux administrateurs de Reddit pour qu'ils l'examinent, l'un n'excluant pas l'autre. Les modérateurs peuvent définir des règles pour leur communauté et les faire appliquer, à condition qu'elles ne contreviennent pas aux conditions d'utilisation et aux autres règles de Reddit. Les modérateurs peuvent configurer AutoModerator, un outil d'aide à la modération des communautés fourni par la plateforme. AutoModerator permet d'effectuer automatiquement certaines tâches, telles qu'envoyer des commentaires utiles à des publications pour rappeler les règles de la communauté aux utilisateurs, ou supprimer ou étiqueter certaines publications par domaine ou mot clé (Reddit Inc., n.d.). Des employés Reddit formés à cet effet sont par ailleurs chargés d'appliquer la politique de contenu à l'ensemble de la plateforme. Enfin, les utilisateurs de Reddit participent eux-mêmes au signalement et au classement des contenus discutables. Ils peuvent les signaler aux modérateurs de la communauté ou aux employés de Reddit. Chaque utilisateur peut aussi voter pour déclasser un élément de contenu. Un fois un certain nombre de votes négatifs atteint, le contenu peut être dégradé ou masqué. Reddit possède des outils internes de création d'empreintes numériques de contenus visant à empêcher la remise en ligne de nouveaux éléments de contenus terroristes déjà identifiés. Reddit bloque également les adresses universelles de domaines connus comme étant contrôlés ou exploités par des organisations identifiées comme étant terroristes. Le coût économique marginal d'une utilisation des outils automatisés pour détecter les contenus discutables est probablement très faible (bien que les coûts fixes puissent être élevés), alors que celui d'un recours à des modérateurs humains pour la même fonction est probablement relativement élevé. Les modérateurs bénévoles n'engendrent aucun coût pour Reddit. Reddit n'est pas membre du GIFCT, mais participe au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions	Le non-respect des conditions d'utilisation ou de la politique de contenu de Reddit peut entraîner le retrait du contenu en infraction et la suspension ou la résiliation définitive du compte de l'utilisateur (selon la gravité de l'incident), de son statut de

d'utilisation ou des règles de la communauté	modérateur ou de sa capacité à accéder aux services de Reddit ou à les utiliser. Les modérateurs doivent également respecter les consignes de modération sous peine d'encourir des conséquences, telles que la suppression de certaines fonctionnalités ou de certains droits associés à leur fonction de modérateur. Enfin, si une communauté ne respecte pas la politique de contenu ou les consignes de modération, elle peut être mise en quarantaine ou bannie, selon l'ampleur ou la gravité des violations.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Oui. Reddit publie des rapports de transparence qui comprennent une partie sur les contenus retirés en raison de la violation des règles propres à une communauté ou de la politique de contenu, qui inclut des dispositions sur les contenus violents. Reddit rend compte des contenus identifiés comme terroristes dans le cadre de la procédure de retrait au sens large. Dans son rapport le plus récent (2020), Reddit a indiqué spécifiquement que sur le nombre total de contenus violents retirés (26 986), 557 retraits concernaient des contenus relevant d'organisations étrangères identifiées comme terroristes (selon la liste établie par le ministère des Affaires étrangères des États-Unis). (Reddit Inc., 2020) Dans son rapport de 2018 (Reddit Inc., 2018), Reddit expliquait que la toute grande majorité (vers les 2/3) des retraits de contenus sur Reddit étaient effectués au sein de sous-reddits individuels (les « communautés ») par les modérateurs de ces communautés. Les suppressions en question résultent essentiellement du non-respect des règles propres aux communautés concernées définies par celles-ci et leurs modérateurs. En tant que telles, elles ne représentent donc pas nécessairement autant de violations de la politique de contenu de la plateforme. Même si l'application de ces règles et la mise en œuvre de la politique de contenu de Reddit peuvent parfois se chevaucher, les mesures prises par les modérateurs restent totalement distinctes de celles des administrateurs de Reddit.
8. Informations ou types de données figurant dans les rapports de transparence	Le rapport indique le nombre total et le pourcentage de contenus retirés par les modérateurs des communautés et les administrateurs de Reddit suite à la violation de la politique de contenu, ainsi que le nombre total et le pourcentage de cas de manipulation du contenu (messages indésirables et autres activités non légitimes) ; le nombre et le pourcentage de violations de la politique de contenu suivies d'un retrait effectué par les modérateurs des communautés et par les administrateurs de Reddit, réparties par catégorie (harcèlement, sexualisation de mineurs, contenu violent, pornographie involontaire, marchandises réglementées, informations privées, usurpation d'identité et évasion de ban), ainsi que le type de contenu visé (image ou photo, vidéo, texte, flux et diffusion en direct, publication croisée) ; le nombre de comptes supprimés et suspendus par les administrateurs de Reddit en raison d'une violation de la politique de contenu ou de manipulation de contenu (messages indésirables) ; le nombre de suppressions de communautés (par suite d'infractions à la politique de contenu

	ou de l'absence de modération) ; le nombre de communautés placées en quarantaine ; le nombre de rapports d'utilisateurs reçus par Reddit pour signaler des violations potentielles de sa politique et le pourcentage de rapports à la suite desquels des mesures ont été prises par les administrateurs de Reddit ; le nombre total de recours reçus par Reddit et leur répartition selon qu'ils ont été admis ou rejetés. Le rapport fait également état des demandes de retrait ou de divulgation d'informations de compte reçues par Reddit d'autorités publiques et de services chargés de l'application des lois, par pays, en indiquant s'il a été donné suite à ces demandes ou non. Il mentionne encore d'autres types de demandes de suppression émanant d'acteurs privés (tels que des avocats), également réparties par pays, et leur statut, à savoir si une suite leur a été réservée. Le rapport comprend en outre un certain nombre de listes détaillées de demandes de suppression au titre du droit d'auteur et de copie en vigueur ainsi que de mesures prises dans le cadre de la Digital Millennium Copyright Act aux États-Unis.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Informations non communiquées.
10. Fréquence de publication des rapports de transparence	La fréquence de publication est annuelle.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. La vidéo de l'attentat de Christchurch a été mise en ligne sur l'une des communautés de Reddit. (Hatmaker, 2019) En conséquence, les administrateurs de la plateforme ont exclu toute la communauté de Reddit. Voir également la section 7, plus haut.

15. Kuaishou

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition particulière des contenus terroristes et extrémistes violents. Les conditions d'utilisation de Kuaishou interdisent toutefois aux utilisateurs de mettre en ligne, télécharger, envoyer ou diffuser des informations qui enfreignent la législation chinoise, cela comprenant les contenus incitant à la haine ou à la discrimination ethnique ainsi que ceux favorisant la violence, l'homicide et le terrorisme.

2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.kuaishou.com/about/policy.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Kuaishou indique qu'il est en droit de contrôler et de vérifier les contenus mis en ligne ou publiés par les utilisateurs en vertu des exigences des autorités publiques et qu'il a le droit de réserver aux contenus un traitement selon les lois et règlements applicables.
4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.
4.2 Mécanismes de recours en cas de suppression ou	Un recours peut être introduit pour le cas où un compte aurait été supprimé par erreur. Les instructions à suivre sont disponibles sur https://www.kuaishou.com/help/feedback/2664?categoryId=hot.

d'autres décisions de sanction	
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Kuaishou a mis en place un système permettant aux utilisateurs de signaler les contenus illicites ou répréhensibles. Les signalements sont ensuite vérifiés et traités par des modérateurs. Le coût économique marginal d'un recours à des modérateurs humains pour détecter les contenus répréhensibles est probablement relativement élevé. Kuaishou n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	En cas d'infraction à ses conditions d'utilisation, Kuaishou peut restreindre ou interdire l'utilisation de la plateforme et des services liés, fermer ou désactiver le compte de l'utilisateur concerné et, le cas échéant, contacter les autorités compétentes.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les	Sans objet.

informations et données figurant dans les rapports de transparence	
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Information inconnue.

16. Telegram

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. Les conditions d'utilisation de Telegram interdisent toutefois la promotion de la violence sur ses chaînes publiques. Cette interdiction ne s'applique pas aux « échanges secrets ».
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://telegram.org/tos.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures	Aucune procédure n'est indiquée. Telegram stipule que s'il reçoit une ordonnance judiciaire confirmant qu'un utilisateur est suspecté de terrorisme, il peut communiquer l'adresse IP et le numéro de téléphone de cet utilisateur aux autorités concernées. Il précise que cela ne s'est encore jamais produit (Telegram, n.d.).

de recours contre ces décisions	
4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Aucune procédure de recours n'est indiquée.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Telegram autorise les utilisateurs à signaler les contenus qui ne respectent pas ses règles. Il dispose aussi d'une équipe qui examine les contenus des chaînes publiques. En 2016, Telegram a créé la chaîne « ISIS Watch » pour montrer les efforts qu'il déploie pour supprimer les contenus encourageant le terrorisme sur ses chaînes et bots publics. D'après la chaîne, Telegram aurait supprimé plus de 200 000 chaînes et bots publics en lien avec l'EIIL (Telegram, n.d.). Le coût économique marginal d'un recours à des modérateurs humains pour détecter les contenus problématiques est probablement relativement élevé. Telegram n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	Aucune notification n'est indiquée.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.

10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. Plusieurs attaques terroristes ont été coordonnées sur Telegram (Bennett, 2019) (Hayden, Far-Right Extremists Are Calling for Terrorism on the Messaging App Telegram, 2019) (Bennett, 2019) (Hayden, Far-Right Extremists Are Calling for Terrorism on the Messaging App Telegram, 2019).

17. Snapchat

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. Cependant, dans les règles communautaires de Snapchat, Snap indique au point Terrorisme qu'il est interdit aux organisations terroristes d'utiliser sa plateforme et qu'il n'a aucune tolérance pour les contenus qui prônent ou font la promotion du terrorisme. Les règles que Snap rend publiques ne définissent pas le terme « organisation terroriste », mais, en interne, Snap applique la définition suivante : « Une **organisation terroriste étrangère** est une organisation désignée comme telle par le ministère des Affaires étrangères des États-Unis .Le **terrorisme** est défini comme le recours illicite à la violence et à l'intimidation, en particulier contre des civils, pour atteindre des objectifs politiques. » Snap interdit également tout contenu qui encourage la discrimination ou la violence sur la base de la race, de l'appartenance ethnique, de l'origine nationale, de la religion, de l'orientation sexuelle, de l'identité de genre, du handicap ou du statut d'ancien combattant.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.snap.com/en-GB/terms/#terms-row et https://www.snap.com/en-GB/community-guidelines.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Sans objet. Snapchat ne propose pas la diffusion de contenu en direct.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier,	Snap indique qu'il se réserve le droit de supprimer tout contenu (i) qui enfreint selon lui ses conditions d'utilisation ou ses règles communautaires ou (ii) si une telle suppression est nécessaire pour satisfaire ses obligations légales.

existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Snap précise qu'il soutient les Santa Clara Principles on Transparency and Accountability in Content Moderation (Santa Clara University's High Tech Law Institute, n.d.), selon lesquels les entreprises doivent informer les utilisateurs dont le contenu est retiré ou dont le compte est suspendu des raisons de ce retrait ou de cette suspension. Les principes disposent également que les entreprises doivent donner la possibilité de faire appel des suppressions de contenu et des suspensions de compte. Les règles de Snapchat ne prévoient toutefois pas pour l'instant l'envoi de notifications en cas de suppression de contenu ni de procédures d'appel des décisions de suppression de contenu ou de suspension de compte.
4.1 Notifications des suppressions ou des autres décisions de sanction	Le retrait par l'équipe Trust & Safety d'un snap s'accompagne de l'envoi au titulaire du compte d'un avertissement contenant un lien vers les règles communautaires de Snap. Lorsque l'équipe prend des mesures à l'encontre d'un compte, son titulaire est informé de la résiliation pour violation des règles communautaires ou des conditions d'utilisation de Snap.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	En fonction de l'infraction, un utilisateur peut se voir empêché de créer de nouveaux comptes pendant une période de six mois ou plus. Pour faire appel d'une telle décision, l'utilisateur peut s'adresser à Snap par le biais de son site d'assistance.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Les utilisateurs peuvent signaler les contenus contraires aux règles de Snapchat (Snap Inc., n.d.). Platform Integrity évalue les contenus susceptibles de non-respect sur la base des rapports des utilisateurs, des notifications des signaleurs de confiance, des remontées internes et des outils de détection automatique. Snap reçoit les bulletins de l'US National Counter Terrorism Center, des alertes d'Europol et des services chargés de l'application des lois ainsi que des rapports de fournisseurs tiers pour les contenus potentiellement extrémistes publiés sur Snapchat. Snap dispose d'une équipe spécialisée dans la sécurité qui travaille 24 heures sur 24. Les contenus qui enfreignent les règles de Snapchat sont retirés. Le coût économique marginal d'un recours à des modérateurs humains pour détecter les contenus répréhensibles est probablement relativement élevé. Snapchat n'est pas membre du GIFCT, mais participe au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.

6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	Les clichés qui enfreignent les règles de Snap sont retirés. De même, les comptes enfreignant de manière répétée les règles ainsi que ceux visant à diffuser des contenus en infraction sont supprimés de la plateforme. Les utilisateurs violant plusieurs fois ou massivement les règles de Snap peuvent se voir empêchés de créer de nouveaux comptes pendant six mois ou plus. D'une manière générale, si un utilisateur enfreint ses conditions d'utilisation ou ses règles communautaires, Snapchat peut supprimer le contenu inapproprié, résilier le compte concerné et avertir les autorités. Si un compte est résilié parce qu'il a enfreint les règles de Snapchat, son propriétaire ne peut plus utiliser Snapchat.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non. Snap publie toutefois des rapports de transparence (Snap Inc., 2015-2020) dont une section porte sur les demandes de suppression de contenu émanant des autorités publiques pour signaler des cas de violation de ses conditions d'utilisation ou règles communautaires. Celles-ci mentionnent notamment que les contenus terroristes et extrémistes violents sont interdits. Pour autant, ces rapports n'incluent pas d'indicateurs concernant spécifiquement les contenus terroristes et extrémistes violents.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Snap publie des rapports de transparence à raison de deux par an.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. Par exemple, la vidéo de l'attentat terroriste de Nice, en 2016, a été diffusée dans les stories de Snapchat et apparaissait dans les recherches effectuées avec la fonction Explorer (Manileve, 2016).

18. Pinterest

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition particulière des contenus terroristes et extrémistes violents. Les règles communautaires de Pinterest énoncées dans son *Guide pour la communauté* stipulent que les organisations et personnes dangereuses sont interdites sur Pinterest. Elles font ainsi référence aux groupes qui encouragent, louent, promeuvent ou aident des acteurs ou groupes dangereux et leurs activités, dont : • les extrémistes ; • les organisations terroristes ; • les gangs et autres organisations criminelles. Les termes ci-dessus ne sont pas définis. Par ailleurs, Pinterest interdit les contenus haineux ou les personnes et groupes qui promeuvent des activités haineuses, cela comprenant : • les insultes ou les stéréotypes négatifs, les caricatures et les généralisations ; • le soutien aux groupes haineux et aux personnes promouvant des activités haineuses, des préjugés et des théories du complot ; • l'approbation ou la banalisation de la violence en raison de l'appartenance d'une victime à un groupe vulnérable ou protégé ; • le soutien à la suprématie blanche, à la limitation des droits des femmes et à d'autres idées discriminatoires ; • les théories du complot fondées sur la haine et la mésinformation, comme la négation de l'Holocauste ; • le déni de l'identité de genre ou de l'orientation sexuelle d'une personne, et le soutien à la thérapie de conversion et aux programmes connexes ; • les attaques d'individus, y compris de personnalités publiques, en raison de leur appartenance à un groupe vulnérable ou protégé ; • la moquerie ou l'attaque au sujet des croyances, des symboles sacrés, des mouvements ou des institutions des groupes protégés ou vulnérables identifiés ci-dessous. Les groupes protégés et vulnérables comprennent : les personnes regroupées en fonction de leur race, couleur, caste, ethnie, statut d'immigration, origine nationale, religion ou foi, sexe ou identité de genre, orientation sexuelle, handicap ou état de santé, réels ou perçus. Ils comprennent également les personnes regroupées en fonction de leur statut socio-économique inférieur, âge, poids ou taille, grossesse ou statut d'ancien combattant.

2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://policy.pinterest.com/en-gb/terms-of-service et https://policy.pinterest.com/en-gb/community-guidelines.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Sans objet. Pinterest ne propose pas la diffusion de contenu en direct.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Pinterest indique se réserver le droit de supprimer ou de modifier les contenus des utilisateurs ou de modifier la manière dont ils sont utilisés sur la plateforme, pour quelque raison que ce soit. Cela concerne aussi les contenus des utilisateurs qui ne respectent pas ses règles. Dans son *guide pour la communauté*, Pinterest fait remarquer qu'il restreint la diffusion des contenus ou retire les contenus et les comptes en cas d'infractions à ses règles relatives aux activités haineuses et aux organisations et personnes dangereuses.
4.1 Notifications des suppressions ou des autres décisions de sanction	Pinterest avertit « dans la plupart des cas » les utilisateurs de la suppression de leur contenu, mais il n'est pas précisé dans quels cas des notifications sont effectivement envoyées.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Il n'existe pas de mécanisme de recours en cas de retrait de contenu, mais il est possible de faire appel des décisions de suspension de compte (Pinterest, n.d.).
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Pinterest a mis en place un système permettant aux utilisateurs de signaler les contenus qui enfreignent ses règles. Il dispose d'une équipe de modérateurs qui surveillent les contenus. Les contenus terroristes et violents sont supprimés lorsqu'ils sont détectés. Pinterest indique qu'il collabore avec les entreprises du secteur, les autorités et les experts en sécurité pour identifier les groupes terroristes. Le coût économique marginal d'un recours à des modérateurs humains pour détecter les contenus répréhensibles est probablement relativement élevé.

	Pinterest est membre du GIFCT, mais ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	En cas de violation de ses règles, Pinterest peut résilier ou suspendre l'accès de l'utilisateur concerné à la plateforme immédiatement et sans préavis. Les notifications relatives à ces mesures sont envoyées à la discrétion de Pinterest.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non. Pinterest publie des rapports de transparence (Pinterest, 2014-2020) dont une section porte sur les demandes de suppression de contenu des autorités et de tiers privés signalant des violations de ses conditions d'utilisation ou de la législation locale, mais qui ne mentionnent pas précisément les retraits des contenus terroristes et extrémistes violents.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Information inconnue.

19. Twitter

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition particulière des *contenus* terroristes et extrémistes violents, mais la plateforme possède une politique relative au terrorisme et à l'extrémisme violent explicitant ce que Twitter considère comme une organisation terroriste ou extrémiste violente et fournissant des exemples de contenus enfreignant ses règles.
	La section Sécurité des *règles de Twitter* interdit explicitement le terrorisme et l'extrémisme violent.

	D'après la politique relative au terrorisme et à l'extrémisme violent de Twitter, les utilisateurs ne peuvent pas proférer de menaces de terrorisme ou d'extrémisme violent, ni les promouvoir. Twitter affirme qu'il n'accepte pas les organisations terroristes ou les personnes ou les groupes extrémistes violents qui leur sont associés et encouragent leurs activités illégales. Il fonde son appréciation de ces organisations sur les définitions du terrorisme national et international, sans toutefois préciser ces dernières. Les organisations sont aussi évaluées en fonction de ses propres critères de définition des groupes extrémistes. Les organisations qui : • sont identifiées par le biais de leur objectif affiché, de leurs publications ou de leurs agissements comme un groupe extrémiste ; • ont commis ou commettent des actes de violence ou encourageant la violence pour soutenir leur cause ; • visent des civils dans leurs actes ou pour promouvoir la violence, sont considérées comme des groupes extrémistes violents. Twittter examine les activités d'un groupe sur sa propre plateforme et à l'extérieur de celle-ci pour déterminer s'il commet ou encourage des actes de violence à l'encontre de civils pour servir une cause politique, religieuse ou sociale. Il fournit les exemples de contenus contraires à sa politique sur le terrorisme et l'extrémisme violent suivants : • commettre ou promouvoir des actes au nom d'une organisation terroriste ou d'un groupe extrémiste violent ; • recruter pour une organisation terroriste ou un groupe extrémiste violent ; • fournir ou distribuer des services (par exemple financiers, de média/propagande) pour servir les objectifs affichés d'une organisation terroriste ou d'un groupe extrémiste violent ; • utiliser les insignes ou les symboles d'organisations terroristes ou de groupes extrémistes violents pour les promouvoir. Le paragraphe relatif à la *conduite haineuse* des règles de Twitter indique que les utilisateurs ne peuvent pas menacer d'autres personnes, les harceler et inciter à la violence envers elles sur la base de critères de race, d'origine ethnique, de nationalité, d'orientation sexuelle, de sexe, d'identité sexuelle, d'appartenance religieuse, d'âge, de handicap ou de maladie grave. Les comptes qui ont pour principal objectif d'inciter à nuire en fonction de ces critères sont interdits. Les utilisateurs ne peuvent pas non plus publier d'images ou de symboles de haine pour leur bannière ou leur photo de profil, ni employer leur nom d'utilisateur, leur identifiant ou leur biographie pour se livrer à un comportement inapproprié, tel

	que le harcèlement ciblé ou l'envoi de messages de haines à l'encontre d'une personne, d'un groupe ou d'une catégorie précise. Cette politique interdit les menaces de violence, les messages de souhait ou d'espoir qu'une personne ou un groupe de personnes subisse un préjudice grave ou appelant à causer un préjudice grave pour celles-ci, les références à des meurtres de masse, à des événements violents ou à des actes de violence spécifiques dont des groupes précis ont été les principales cibles ou victimes, et les messages suscitant la méfiance envers d'un groupe spécifique faisant l'objet d'une protection. Sont également interdits les insultes, qualificatifs et clichés racistes et sexistes, répétés ou sans consentement de la part de la personne visée, ou tout autre contenu dégradant une personne, les imageries haineuses (tels que les logos, symboles ou images dont le but est d'encourager l'hostilité et la méchanceté à l'encontre d'autres personnes sur la base de leur race, religion, handicap, orientation sexuelle, identité sexuelle ou origine ethnique/nationale). Enfin, la politique relative à l'apologie de la violence interdit l'apologie de la violence, en particulier les événements violents où des personnes sont visées sur la base de caractéristiques faisant l'objet d'une protection (telles que la race, l'origine ethnique, la nationalité, l'orientation sexuelle, le sexe, l'identité sexuelle, l'appartenance religieuse, l'âge, le handicap ou la maladie grave), ce comportement pouvant inciter à la violence motivée par la haine ou l'intolérance ou l'exacerber. En vertu de cette politique, les utilisateurs ne peuvent pas faire l'apologie de crimes violents ou d'événements violents visant des personnes précises en raison de leur appartenance à un groupe spécifique, ni les auteurs de tels actes, ni les célébrer, les louer ou les cautionner. L'apologie est définie comme comprenant les déclarations élogieuses, ou qui célèbrent ou approuvent des actions, telles que « Je suis heureux que cela se soit produit », « Cette personne est mon héros », « J'aimerais que plus de gens fassent des choses comme ça » ou « J'espère que cela incitera d'autres personnes à passer à l'acte ». Les infractions à cette politique comprennent notamment le fait de glorifier, de louer, d'approuver ou de célébrer : • les actes violents commis par des civils et ayant entraîné la mort ou des blessures graves, comme les meurtres ou fusillades de masse ; • les attaques perpétrées par des organisations terroristes ou des groupes extrémistes violents ; • les événements violents qui ont ciblé des groupes protégés, comme l'Holocauste ou le génocide rwandais.
2. Manière dont les conditions d'utilisation ou les	Les textes sont disponibles sur https://help.twitter.com/en/rules-and-policies/twitter-rules, https://help.twitter.com/en/rules-and-policies/violent-groups, https://help.twitter.com/en/rules-and-policies/hateful-

règles de la communauté sont communiquées	conduct-policy et https://help.twitter.com/en/rules-and-policies/glorification-of-violence.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Twitter possède différentes mesures de sanction qu'il peut appliquer en cas de violation de ses règles (Twitter, n.d.). a. *Application au niveau du tweet* : s'applique au contenu qui enfreint les politiques de Twitter, mais la plateforme pense qu'il est dans l'intérêt du public que ce contenu demeure accessible. Le tweet est masqué par un avis donnant la possibilité à l'utilisateur de le voir ou non. Les tweets d'intérêt public n'apparaissent pas dans les fils d'actualités Tweets populaires, dans la recherche sans risque, dans les recommandations push et via l'onglet Notifications, dans les recommandations par courriel et SMS, dans le fil d'événements en direct et dans l'onglet Explorer. Twitter intervient au niveau du tweet pour ne pas prendre de mesures trop sévères à l'encontre d'un compte par ailleurs irréprochable ayant commis une erreur et enfreint ses règles. Les mesures prises à ce niveau peuvent consister à limiter la visibilité du tweet, exiger son retrait et le masquer jusqu'à ce qu'il soit retiré. b. *Application au niveau des messages privés* : lors d'une conversation par messages privés, quand un participant signale l'autre, Twitter empêche la personne en infraction d'envoyer des messages à la personne qui l'a signalée. La conversation est aussi supprimée de la boîte de réception de la personne ayant effectué le signalement. Dans une conversation de groupe par messages privés, le message privé en infraction peut être placé derrière un avis pour que personne d'autre dans le groupe ne puisse le voir à nouveau. c. *Application au niveau du compte* : Twitter agit au niveau du compte s'il détermine qu'une personne a enfreint ses règles d'une manière particulièrement flagrante ou les a enfreintes à maintes reprises, même après avoir reçu des notifications de sa part. Les mesures prises peuvent être les suivantes :

	- Demande de modifications du média ou du profil : si le profil ou un contenu média d'un compte n'est pas conforme aux politiques de Twitter, ce dernier peut le rendre provisoirement indisponible et demander au contrevenant de modifier le média ou les informations de son profil pour se mettre en conformité. Twitter indique également la politique enfreinte par le profil ou le média. - Placement d'un compte en mode lecture seule : si un compte par ailleurs irréprochable semble traverser un épisode riche en comportements inappropriés, Twitter peut le placer provisoirement en lecture seule, ce qui limite sa capacité à tweeter, retweeter ou aimer du contenu jusqu'à ce que l'utilisateur concerné ait retrouvé son sang-froid. La personne concernée pourra lire ses fils et sera uniquement en mesure d'envoyer des messages privés à ses abonnés. Lorsqu'un compte est en mode lecture seule, les autres utilisateurs sont toujours en mesure de le voir et d'interagir avec lui. La durée de cette sanction peut aller de 12 heures à sept jours, selon la nature de l'infraction. - Vérification de la propriété d'un compte : afin de s'assurer qu'aucun contrevenant n'abuse de l'anonymat que Twitter offre pour harceler d'autres personnes sur la plateforme, le service peut demander au détenteur du compte concerné d'en confirmer la propriété avec un numéro de téléphone ou une adresse électronique. Cette mesure aide également à identifier tout contrevenant gérant plusieurs comptes à des fins abusives et à prendre des mesures à l'encontre de tels comptes. Si un compte a été verrouillé dans l'attente de mesures correctives (fournir un numéro de téléphone, par exemple), il disparaît du nombre d'abonnés, des retweets et des « J'aime » jusqu'à ce que la situation soit normalisée. - Suspension définitive : sanction la plus sévère. La suspension définitive d'un compte empêche la consultation du compte à l'échelle mondiale, et le contrevenant ne peut créer aucun nouveau compte. Lorsqu'il examine s'il convient d'appliquer des sanctions, Twitter prend en considération différents critères, à savoir, notamment, si : • le comportement prend pour cible une personne, un groupe ou une catégorie de personnes protégée ;

	• le signalement a été introduit par la cible du comportement abusif ou par une personne à proximité ; • l'utilisateur a déjà enfreint les règles plusieurs fois dans le passé ; • l'infraction présente un degré de gravité élevé ; • le contenu est un sujet d'intérêt public (Twitter, n.d.).
4.1 Notifications des suppressions ou des autres décisions de sanction	Des notifications sont envoyées lorsque Twitter demande à un utilisateur de modifier son comportement pour respecter ses règles (demande de modification des contenus ou du profil) ou en cas de suspension définitive d'un compte. Dans ce dernier cas, Twitter informe l'utilisateur que son compte a été suspendu en raison d'infractions et indique la ou les politiques il a enfreintes ainsi que le contenu incriminé.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Les utilisateurs peuvent faire appel des suspensions définitives s'ils estiment que Twitter a commis une erreur. Si la validité de la suspension est confirmée à l'issue de la procédure d'appel, Twitter fournit des informations relatives à la politique enfreinte par le compte.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Twitter dispose de trois moyens de détecter les contenus contraires à ses règles. 1. Signalement par les utilisateurs Twitter encourage les utilisateurs à signaler les infractions à ses règles. Des modérateurs examinent les signalements pour déterminer si le contenu concerné enfreint effectivement des règles. Twitter possède une équipe mondiale chargée de l'application de ses règles 24 heures sur 24 dans toutes les langues prises en charge par la plateforme. 2. Détections proactives des contenus Twitter utilise également des outils développés en interne pour repérer les infractions à ses règles, telles que la mise en ligne de contenus terroristes et extrémistes violents, à partir des contenus publiés, tels que des vidéos notoirement créées par des organisations terroristes. 3. Détections proactives des comportements Twitter utilise des outils développés en interne pour repérer les infractions à ses règles, telles que la mise en ligne de contenus terroristes et extrémistes violents, à partir des comportements des utilisateurs qui peuvent être associés à ceux d'organisations terroristes. Twitter envisage de développer sa propre technologie antispam pour repérer de manière proactive les contenus terroristes et extrémismes violents, les méthodes employées par certains groupes se rapprochant de celles utilisées pour diffuser des messages indésirables.

	Le coût économique marginal d'une utilisation des outils automatisés pour détecter les contenus terroristes et extrémistes violents est probablement très faible (bien que les coûts fixes puissent être élevés), alors que celui d'un recours à des modérateurs humains pour la même fonction est probablement relativement élevé. Twitter est membre du GIFCT et participe au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	Les infractions à la politique sur le terrorisme et l'extrémisme violent entraînent une suspension immédiate et définitive du compte concerné. Les infractions à la politique relative à la conduite haineuse entraînent différentes sanctions, en fonction d'un certain nombre de facteurs, notamment la gravité de l'infraction et les éventuelles infractions commises précédemment par l'utilisateur. Twitter peut par exemple demander à un utilisateur de supprimer le contenu contraire aux règles et placer son compte en mode lecture seule pendant un certain temps. De nouvelles violations entraîneront un placement du compte en mode lecture seule plus long, voire une suspension définitive du compte. Si un utilisateur fait essentiellement preuve d'un comportement inapproprié ou est réputé avoir partagé une menace violente, son compte sera suspendu définitivement dès le premier examen. Les infractions à la politique relative à l'apologie de la violence entraînent des mesures différentes selon leur gravité et les précédentes infractions commises par le compte. Lors de la première infraction à la politique, Twitter demande à l'utilisateur de retirer le contenu incriminé. Il bloque également temporairement l'accès de l'utilisateur à son compte. Si les infractions se poursuivent après l'envoi d'un avertissement, le compte est suspendu définitivement.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Oui. Les rapports de transparence de Twitter (Twitter, 2012-2020) comportent une section consacrée à l'application des règles, ces dernières comprenant les politiques décrites à la section 1 ci-dessus.
8. Informations ou types de données figurant dans les rapports de transparence	Twitter fait état des indicateurs suivants : • Comptes ayant fait l'objet de mesures : nombre de comptes qui ont été suspendus ou dont des contenus ont été retirés à la suite d'une infraction aux règles. • Contenus supprimés : nombre d'éléments de contenu uniques (tels que des tweets ou des photos ou images de profil, des bannières ou des biographies) que Twitter a demandé à des titulaires de compte de retirer au titre qu'ils violaient ses règles ; • Comptes suspendus : nombre de comptes qui ont été

	suspendus à la suite d'une infraction aux règles. Chacun des indicateurs ci-dessus fait l'objet d'une décomposition en fonction des politiques sous-tendant les règles, dont celles visées à la section 1 (c'est-à-dire le terrorisme et l'extrémisme violent, la conduite haineuse et l'apologie de la violence). Pour ce qui concerne spécifiquement le terrorisme et l'extrémisme violent, Twitter rend compte du pourcentage de comptes ayant fait l'objet de mesures à la suite d'une identification proactive. Twitter inclut également dans les données rapportées des tendances, dont certaines concernent les contenus terroristes et extrémistes violents. Ainsi, dans son dernier rapport, Twitter a observé une diminution de 9 % du nombre de comptes pour lesquels des mesures ont dû être prises au titre de la violation de sa politique en matière de terrorisme et d'extrémisme violent par rapport à la période de rapport précédente.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Les « comptes signalés » reflètent le nombre total de comptes que les utilisateurs ont signalés comme ayant potentiellement enfreint les règles de Twitter. Pour obtenir des chiffres fiables, Twitter ne comptabilise qu'une seule fois les comptes signalés plusieurs fois (que plusieurs utilisateurs aient signalé un compte pour la même infraction potentielle ou que plusieurs utilisateurs aient signalé le même compte pour différentes infractions potentielles). Pour ces statistiques, Twitter ne comptabilise également qu'une seule fois les signalements concernant des tweets particuliers : s'il reçoit des signalements concernant plusieurs tweets émis par le même utilisateur, il ne les comptabilise qu'une fois dans les « comptes signalés ». Les « comptes faisant l'objet de mesures » reflètent le nombre de comptes pour lesquels Twitter a pris des sanctions pendant la période couverte par le rapport. Ces mesures sont l'une de celles décrites à la section 4 ci-dessus. Pour obtenir des chiffres fiables, Twitter ne comptabilise qu'une seule fois les comptes qui ont fait l'objet de plusieurs mesures pour la même infraction. S'il a pris des mesures pour un tweet ou un compte au titre de plusieurs politiques, il comptabilisera le compte pour chacune d'entre elles. Cependant, s'il a pris à plusieurs reprises des mesures pour un tweet ou un compte au titre de la même politique (il a par exemple placé provisoirement un compte un mode lecture seule, puis a demandé à l'utilisateur de modifier ses contenus ou son profil à la suite de la même violation), il comptabilisera le compte une seule fois pour la violation de cette politique.
10. Fréquence de publication des rapports de transparence	Les rapports font l'objet d'une édition semestrielle.

11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. Voir les sections 7 et 8 ci-dessus.

20. Douban

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition particulière des contenus terroristes et extrémistes violents. Les conditions d'utilisation de Douban interdisent toutefois aux utilisateurs de mettre en ligne, de diffuser ou d'utiliser de quelque façon que ce soit des contenus contenant de la violence gratuite ou encourageant la violence, le racisme, la discrimination, l'extrémisme religieux, la haine ou les préjudices physiques de quelque nature qu'ils soient contre un groupe ou une personne, ou qui sont répréhensibles de quelque manière que ce soit.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.douban.com/note/732773017/.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Douban indique se réserver le droit (sans en avoir l'obligation) d'examiner à sa seule discrétion les contenus des utilisateurs. Il précise également qu'il peut à son entière discrétion supprimer ou modifier un contenu à tout moment et pour quelque raison que ce soit, avec ou sans envoi d'une notification à l'utilisateur.
4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Aucune procédure de recours n'est indiquée.

5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Aucune information n'est communiquée. Douban n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	En cas de violation des conditions d'utilisation, Douban est autorisé à suspendre les droits de l'utilisateur concerné à utiliser ses services. Il peut également résilier son compte.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Information inconnue.

17. LinkedIn

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Les *politiques de la communauté professionnelle LinkedIn* comportent différentes sections interdisant les contenus terroristes et extrémistes violents : **Ne promouvez pas, ne menacez pas et n'incitez pas à la violence :** il est interdit de menacer de violence ou d'inciter à la violence. Les individus ou les groupes ont interdiction de promouvoir ou de se livrer à de la violence, à des dommages matériels ou à toute activité de criminalité organisée. LinkedIn ne peut être utilisé pour exprimer un soutien à de tels individus ou groupes ou pour glorifier un fait de violence, quel qu'il soit. **Ne partagez pas de contenu nuisible ou choquant :** tout contenu excessivement macabre ou choquant est prohibé. Est ici visé tout contenu sadique ou qui cherche à choquer, tel que la description de violences physiques graves. Les contenus et activités qui promeuvent, organisent, dépeignent ou favorisent les activités criminelles sont interdits. Les contenus décrivant ou promouvant la fabrication d'armes, la toxicomanie et les menaces de vol sont interdits LinkedIn interdit à ses utilisateurs de promouvoir et de se livrer à la publication de contenus à caractère sexuel explicite sans consentement (par exemple, la vengeance pornographique), de services d'escort-girls, de prostitution, d'exploitation d'enfants ou de traite d'êtres humains. Les utilisateurs sont encore tenus de ne pas partager de contenu ou d'activités qui promeuvent ou encouragent le suicide ou tout type de blessures auto-infligées, comme l'automutilation et les troubles alimentaires. S'ils repèrent des signes indiquant qu'une personne envisage peut-être de se faire du mal, les utilisateurs sont priés de le **signaler**. **Ne publiez pas de contenu terroriste ou promouvant le terrorisme :** aucune organisation terroriste ni aucun groupe extrémiste violent n'est autorisé sur LinkedIn. Le service n'accepte pas non plus les individus affiliés à de tels groupes ou organisations pour promouvoir leurs activités. Tout contenu décrivant une activité terroriste conçu pour recruter pour le compte d'organisations terroristes ou qui promeut ou soutient le terrorisme de quelque manière que ce soit, ou profère des menaces de terrorisme, ne saurait être toléré.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.linkedin.com/legal/professional-community-policies.
3. Présence de dispositions précises applicables	Oui. La diffusion de contenus en direct doit respecter les conditions d'utilisation et les politiques de la communauté professionnelle LinkedIn, et s'inscrit en outre dans un cadre très précis. Les membres du réseau qui souhaitent utiliser cette fonctionnalité doivent remplir un formulaire de demande d'accès qui sera examiné selon un ensemble spécifique de

aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	critères. Le formulaire de demande est disponible à l'adresse https://www.linkedin.com/help/linkedin/ask/lv-app. LinkedIn précise encore les lignes de conduite et les bonnes pratiques à suivre pour la diffusion de contenus en direct à l'adresse https://www.linkedin.com/help/linkedin/answer/100225?query=linkedin%20live&hcppcid=search.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	LinkedIn encourage ses utilisateurs à signaler les contenus qui enfreignent ses politiques. Lorsqu'un utilisateur signale un contenu publié par un autre utilisateur, celui-ci n'est pas informé de l'identité de l'auteur du signalement et ce dernier ne voit plus le contenu ou la conversation signalé dans son fil d'actualité ou sa boîte de réception. LinkedIn peut examiner le contenu ou la conversation ayant fait l'objet du signalement pour prendre des mesures supplémentaires, telles que le retrait du contenu voire, en cas d'infractions graves ou commises à plusieurs reprises, la suspension du compte si le contenu est contraire à ses conditions d'utilisation ou à ses politiques. LinkedIn a mis en œuvre des fonctionnalités destinées à améliorer la transparence de ses décisions de modération de contenu, à la fois pour les signaleurs et les auteurs. Une fonctionnalité introduite récemment prévoit une boucle de retour permettant aux signaleurs de recevoir une notification au moment du signalement ainsi que lorsque LinkedIn prend une décision sur le signalement. Pour les auteurs dont un contenu est retiré en raison d'une infraction, la boucle de retour leur permet d'être avertis de la suppression et d'introduire un recours. La fonctionnalité est déployée dans un premier temps aux États-Unis, en France et au Canada, avant un élargissement au reste du monde. (Pour plus d'informations : https://blog.linkedin.com/2020/september/29/new-features-help-keep-it-professional.)
4.1 Notifications des suppressions ou des autres décisions de sanction	La suppression d'un contenu est notifiée tant à l'auteur de celui-ci qu'à la personne à l'origine du signalement.
4.2 Mécanismes de recours en	Si un compte a été restreint ou si du contenu a été supprimé et que l'utilisateur pense qu'il s'agit d'une erreur, il peut faire appel de la décision.

cas de suppression ou d'autres décisions de sanction	L'auteur peut également introduire un recours au moment où il est averti de la suppression de son contenu.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Les utilisateurs peuvent signaler les contenus contraires aux politiques de LinkedIn. Des modérateurs examinent les signalements pour déterminer s'il convient de prendre des mesures. Microsoft, Inc., société mère de LinkedIn, indique qu'elle supprime les contenus terroristes publiés sur les services aux consommateurs qu'elle héberge lorsqu'ils sont portés à sa connaissance par des outils de signalement en ligne (Microsoft, 2016). Le coût économique marginal d'un recours à des modérateurs humains pour détecter les contenus répréhensibles est probablement relativement élevé. LinkedIn est membre du GIFCT. Il exploite la base de données d'empreintes numériques du forum ainsi que d'autres outils de détection et d'analyse automatisés pour repérer les contenus terroristes et extrémistes violents potentiels sur sa plateforme.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	La publication d'un contenu contraire aux conditions d'utilisation ou politiques de LinkedIn peut entraîner le retrait du contenu incriminé et, en cas d'infractions graves ou commises à plusieurs reprises, la suspension du compte concerné.
7. Publication par le service de rapports de transparence sur les contenus	Non, pas à proprement parler. LinkedIn publie des rapports de transparence semestriels (LinkedIn, n.d.[106]) dont une section porte sur les demandes de suppression de contenus des autorités publiques en raison de violations de ses conditions d'utilisation ou de la législation locale, ainsi qu'un rapport sur les retraits de contenus au titre des politiques de la communauté professionnelle. Les contenus terroristes et extrémistes violents entrent dans la catégorie des contenus violents ou explicites, qui « comprend les contenus exprimant une menace ou promouvant le

terroristes et extrémistes violents	terrorisme, la violence ou toute autre activité criminelle de même que les contenus extrêmement violents ou destinés à choquer ou à humilier », recouvrant une réalité plus large que les seuls contenus terroristes et extrémistes violents. Le dernier rapport est disponible à l'adresse https://about.linkedin.com/transparency/community-report.
8. Informations ou types de données figurant dans les rapports de transparence	LinkedIn fait état du nombre total de contenus supprimés ainsi que, en particulier, du nombre de contenus supprimés en leur qualité de « contenus violents ou explicites », qui comprennent les contenus terroristes et extrémistes violents. LinkedIn indique aussi le nombre total de demandes de retrait de contenus des autorités publiques à la suite de violations de ses conditions d'utilisation ou de la législation locale, par pays, ainsi que le pourcentage des demandes qui ont entraîné des mesures de la part de LinkedIn. Il n'existe toutefois pas de données concernant spécifiquement les retraits de contenus terroristes et extrémistes violents.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Des explications d'ordre général sont fournies dans le rapport de la communauté, disponible à l'adresse https://about.linkedin.com/transparency/community-report.
10. Fréquence de publication des rapports de transparence	Des rapports sont publiés tous les six mois.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	C'est possible. Des études ont montré que des extrémistes situés aux États-Unis, qui ne sont pas nécessairement violents, ont utilisé LinkedIn pour promouvoir leurs activités et leurs objectifs (START (National Consortium for the Study of Terrorism and Responses to Terrorism), 2018).

22. Baidu Tieba

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition particulière des contenus terroristes et extrémistes violents. Les conditions générales d'utilisation de Baidu Tieba interdisent toutefois les contenus qui incitent à la haine et à la discrimination ethnique, ainsi que ceux favorisant la violence, le meurtre et le terrorisme.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://gsp0.baidu.com/5aAHeD3nKhI2p27j8IqW0jdnxx1xbK/tb/eula.html.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications	Aucune procédure n'est indiquée.

des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	
4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Aucune procédure de recours n'est indiquée.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Baidu Tieba possède un dispositif qui permet aux utilisateurs de signaler des contenus illicites ou répréhensibles. Les signalements sont ensuite vérifiés et traités par des modérateurs, qui décident en dernier ressort de conserver ou de supprimer les contenus. Le coût économique marginal d'un recours à des modérateurs humains pour détecter les contenus répréhensibles est probablement relativement élevé. Baidu Tieba n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences	S'il estime qu'un utilisateur a enfreint ses conditions d'utilisation, Baidu Tieba peut l'exclure à titre provisoire ou définitif, suspendre ou supprimer

en cas de violation des conditions d'utilisation ou des règles de la communauté	son compte ou lui imposer d'autres sanctions conformément aux réglementations en vigueur.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et	Information inconnue.

extrémistes violents	

23. Skype

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Microsoft est la société mère de Skype. Le contrat de services de Microsoft, qui régit Skype, interdit toute activité nuisible à d'autres personnes, telle que publier des contenus terroristes ou extrémistes violents, tenir des propos haineux ou appeler à la violence contre des tiers. Microsoft indique (Microsoft, 2016) que, dans le cadre de ses services, un contenu terroriste désigne un contenu publié par une organisation figurant sur la liste récapitulative du Conseil de sécurité des Nations unies (Conseil de Sécurité des Nations Unies), ou visant à soutenir une telle organisation, qui représente explicitement la violence, encourage les actes violents, cautionne une organisation terroriste ou ses actes, et incite à rejoindre ces groupes. La liste récapitulative du Conseil de sécurité des Nations Unies répertorie les groupes considérés par le Conseil de sécurité des Nations unies comme des organisations terroristes. Il n'est pas fourni de définition de l'extrémisme violent, mais les conditions d'utilisation de Skype interdisent aux utilisateurs de soumettre ou de publier des contenus haineux, inappropriés, illégaux, racistes, insultants ou répréhensibles de quelque manière que ce soit. Dans son Digital Safety Content Report (Microsoft, 2021), Microsoft explique clairement que les « contenus tant terroristes qu'extrémistes violents sont interdits sur les plateformes et services Microsoft » et que le code de conduite contractuel des services Microsoft (Microsoft Services Agreement Code of Conduct) interdit la « publication de contenu terroriste ou extrémiste violent ».
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Le contrat de services Microsoft est disponible sur https://www.microsoft.com/en-us/servicesagreement. Voir aussi https://www.skype.com/en/legal/ios/tos/#1.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.

4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Skype dispose d'une procédure de notification et de retrait. S'il reçoit une notification indiquant qu'un contenu publié, téléchargé, édité, hébergé, partagé ou publié sur Skype (hors communications privées) est inapproprié, viole les droits d'un tiers, ou s'il souhaite supprimer ce contenu pour quelque raison que ce soit, il se réserve le droit de le supprimer automatiquement pour quelque raison que ce soit, immédiatement ou dans des délais autres qui peuvent être définis à sa seule discrétion. Ainsi que le prévoit le contrat de services Microsoft, « si vous enfreignez les présentes conditions, nous pouvons (...) cesser de vous fournir les services ou fermer votre compte Microsoft. Nous nous réservons également le droit de supprimer ou de bloquer votre contenu des services à tout moment si nous pensons qu'il pourrait enfreindre la réglementation applicable ou les présentes conditions. Lors des enquêtes relatives aux infractions suspectées des présentes conditions, Microsoft se réserve le droit de consulter votre contenu afin de résoudre le problème. »
4.1 Notifications des suppressions ou des autres décisions de sanction	Les notifications sont envoyées à la discrétion de Microsoft. Selon le contrat de services Microsoft, « si une information doit vous être communiquée concernant un service que vous utilisez, nous vous enverrons les notifications de service (...). Si vous nous avez donné votre adresse e-mail ou votre numéro de téléphone dans le cadre de votre compte Microsoft, vous êtes susceptible de recevoir des notifications de service par e-mail ou SMS, y compris pour vérifier votre identité avant d'enregistrer votre numéro de téléphone mobile et de vérifier vos achats. Vous êtes susceptible de recevoir des notifications de service par d'autres moyens (par exemple, par des messages intégrés au produit). »
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Le formulaire pour faire appel de la suspension d'un compte Microsoft est disponible à l'adresse https://www.microsoft.com/en-us/concern/AccountReinstatement.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Microsoft indique que le code de conduite du contrat de services Microsoft interdit la « publication de contenu terroriste ou extrémiste violent ». Microsoft encourage le signalement de contenu publié par une organisation terroriste, ou visant à soutenir une telle organisation, qui représente explicitement la violence, encourage les actes violents, cautionne une organisation terroriste ou ses actes, et incite à rejoindre ces groupes. Microsoft évalue ces rapports, prend les mesures nécessaires concernant les contenus et, le cas échéant, suspend les comptes associés à des infractions à son code de conduite. En outre, Microsoft met en œuvre différents outils pour détecter les contenus terroristes et extrémistes violents, dont une technologie de comparaison d'empreintes numériques et d'autres formes de détection proactive.

	Lorsque des utilisateurs soumettent des signalements, des modérateurs les évaluent pour déterminer s'il convient de prendre des mesures. Microsoft indique qu'il supprime les contenus terroristes publiés sur les services aux consommateurs qu'il héberge lorsqu'ils sont portés à sa connaissance par des outils de signalement en ligne (Microsoft, 2016). Microsoft recourt à des outils d'analyse et de reconnaissance (telles que PhotoDNA ou MD5) ainsi qu'à d'autres solutions faisant appel à l'intelligence artificielle, notamment des programmes de catégorisation de textes et d'images ainsi que des techniques de détection de toilettage pour détecter les contenus terroristes et extrémistes violents. (Microsoft, 2021) Le coût économique marginal d'un recours à des modérateurs humains pour détecter les contenus répréhensibles est probablement relativement élevé. Microsoft est un membre fondateur du GIFCT et participe au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	La publication d'un contenu contraire aux conditions d'utilisation ou d'autres règles de Skype peut entraîner la résiliation ou la suspension du compte concerné et des services fournis par Skype. Voir également les informations des sections 4 et 4.1 ci-dessus.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Oui. Les chiffres concernant le nombre de contenus terroristes et extrémistes violents pour Skype figurent dans le Digital Safety Content Report de Microsoft (Microsoft, 2021). Ce rapport couvre les produits et services de Microsoft destinés au public, notamment OneDrive, Outlook, Skype, Bing et Xbox. Il convient de noter que les chiffres relatifs aux contenus terroristes et extrémistes violents sont rapportés collectivement pour tous les produits et services de Microsoft destinés au public et non par produit.
8. Informations ou types de données figurant dans les rapports de transparence	• Nombre de contenus terroristes et extrémistes violents ayant fait l'objet de mesures • Nombre de comptes suspendus en raison de contenus terroristes et extrémistes violents • Pourcentage de contenus terroristes et extrémistes violents détectés par Microsoft • Pourcentage de comptes suspendus en raison de contenus terroristes et extrémistes violents qui ont été rétablis après recours
9. Méthodologies appliquées pour	Le terme « contenu ayant fait l'objet de mesures » (*content actioned*) désigne un élément de contenu publié par un utilisateur que Microsoft a retiré de ses produits et services ou

déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	que Microsoft a bloqué pour empêcher les utilisateurs d'y accéder. Le terme « suspension de compte » (*account suspension*) signifie retirer à l'utilisateur la possibilité d'accéder au compte du service de manière soit permanente, soit temporaire. Le terme « détection proactive » (*proactive detection*) indique que le signalement d'un contenu sur les produits ou services est le fait de Microsoft, que ce soit de manière automatisée ou par évaluation manuelle.
10. Fréquence de publication des rapports de transparence	Cette information n'est pas mentionnée.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	C'est possible. Les recherches menées par le Counter Extremism Project ont montré que des individus ont consulté et diffusé des contenus de propagande extrémistes officiels sur Skype (même s'il ne s'agit pas expressément de contenus extrémistes violents) (Counter Terrorism Project).

24. Quora

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. Cependant, dans la norme « Soyez courtois, soyez respectueux », Quora précise au paragraphe « Pas de glorification ou de promotion de la violence » qu'il « peut également interdire et supprimer tout contenu de tout utilisateur qui est un membre confirmé et/ou déclaré d'un groupe figurant sur la liste des organisations étrangères considérées comme terroristes par le département d'État des États-Unis ou dont la participation à des actes de violence de masse ou à des crimes haineux est attestée ».
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.quora.com/about/tos, https://www.quora.com/about/acceptable_use et https://www.quora.com/What-is-Quoras-Be-Nice-Be-Respectful-policy/answer/Quora-Official-Account
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.

4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Quora indique qu'il a le droit, mais pas l'obligation de refuser de distribuer un contenu sur sa plateforme ou de supprimer du contenu. Une violation des règles peut entraîner l'envoi d'un avertissement et, si l'utilisateur persiste, il peut lui être interdit de poser des questions et d'envoyer des réponses et des commentaires (il est bloqué) ou il peut être exclu. (Quora, n.d.) Les blocages et les exclusions peuvent être provisoires. Un utilisateur bloqué ou exclu peut revenir s'il décide de changer de comportement. Le blocage est généralement levé lorsque l'utilisateur demande par message privé le déblocage de son compte.
4.1 Notifications des suppressions ou des autres décisions de sanction	Il n'existe pas de notification de suppression, mais des avertissements en cas de contenu inapproprié, comme indiqué ci-dessus.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Si un utilisateur pense qu'il a été bloqué ou exclu par erreur, il peut faire appel de cette décision.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Les utilisateurs peuvent signaler les contenus qu'ils estiment contraires aux règles de Quora. Les signalements sont transmis à l'équipe de modération de Quora pour examen. Le coût économique marginal d'un recours à des modérateurs humains pour détecter les contenus répréhensibles est probablement relativement élevé. Quora n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	Les contenus contraires à la norme « Soyez courtois, soyez respectueux » peuvent être signalés aux administrateurs, qui peuvent ensuite les supprimer, et les infractions à cette norme peuvent entraîner l'envoi d'un avertissement, le blocage des commentaires, ou le blocage ou l'exclusion de l'utilisateur concerné (voir la section 4 ci-dessus). Selon la gravité de l'infraction, l'utilisateur peut être exclu immédiatement (c'est-à-dire sans avertissement ni blocage préalables). Quora peut aussi résilier ou suspendre le compte d'un utilisateur qui a enfreint l'une de ses règles.

7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. Des questions sur la façon de rejoindre une organisation terroriste ont été posées sur Quora (Lange, 2017).

25. Xigua

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition particulière des contenus terroristes et extrémistes violents. Les conditions d'utilisation de Xigua interdisent toutefois la promotion du terrorisme et de l'extrémisme.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.ixigua.com/user_agreement/.
3. Présence de dispositions précises	Non.

applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Aucune procédure n'est indiquée.
4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Aucune procédure de recours n'est indiquée.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Les utilisateurs peuvent signaler les activités ou les contenus illicites. Xigua n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	La violation des conditions d'utilisation peut entraîner la clôture du compte de l'utilisateur concerné ou la résiliation de son accès aux services de Xigua sans avertissement préalable.

7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Information inconnue.

26. Viber

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition particulière des contenus terroristes et extrémistes violents. La politique de contenu public de Viber indique toutefois que les contenus excessivement violents, notamment ceux qui font l'apologie ou incitent à la violence, sont interdits. Ces contenus recouvrent les représentations ou les descriptions de la violence et les menaces crédibles de violence à l'encontre d'une personne ou d'un groupe. Viber interdit la planification ou la promotion d'actes violents qui pourraient causer directement ou indirectement des préjudices physiques ou psychologiques aux autres.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.viber.com/terms/viber-terms-use/ et https://www.viber.com/terms/viber-public-content-policy/.

3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Sans objet. À l'heure actuelle, Viber ne propose pas de fonction de diffusion de contenu en direct.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Viber indique que lorsqu'il crée une communauté, l'utilisateur en devient automatiquement l'administrateur « superadmin ». Les administrateurs doivent veiller à ce que les contenus envoyés et affichés dans leur compte public ou leur communauté respectent les politiques et conditions d'utilisation de Viber, ainsi que toutes les lois et réglementations en vigueur. Ils ne doivent pas avoir un comportement interdit par l'un de ces textes ni autoriser un tiers à le faire. Les administrateurs peuvent supprimer eux-mêmes des contenus. Viber peut retirer tout ou partie des contenus s'il estime qu'ils sont interdits, illicites ou contraires à ses règles.
4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Les utilisateurs peuvent contacter le service d'assistance de Viber pour faire appel d'un retrait de contenu d'un blocage. Viber examine toutes les demandes qui lui sont adressées.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes	Les utilisateurs peuvent signaler les contenus contraires aux règles de Viber. Viber examine les signalements et fait intervenir une équipe de modérateurs pour décider des mesures les plus appropriées. Viber possède également des algorithmes internes qu'il applique pour détecter certains contenus illicites. Les administrateurs ont la possibilité de supprimer les contenus en infraction de leurs comptes et communautés. Il est difficile de déterminer dans quelle mesure les contenus publiés sur Viber sont modérés. Ses conditions d'utilisation stipulent que Viber ne s'engage pas à surveiller les dialogues en ligne publics ou autres forums et que les contenus qui y sont publiés ne relèvent pas de sa responsabilité. En outre, ses

numériques ou d'adresses URL)	principales fonctionnalités étant chiffrées, il s'avère impossible de modérer les contenus qui sont diffusés à travers celles-ci. Les fonctionnalités publiques, telles que les communautés ou les dialogues en ligne, ne sont toutefois pas chiffrées de bout en bout et Viber peut en examiner les contenus à la suite d'un signalement et les retirer, le cas échéant. Le coût économique marginal d'un recours à des modérateurs humains pour détecter les contenus répréhensibles est probablement relativement élevé. Les utilisateurs modérateurs n'engendrent aucun coût pour Viber. Viber n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	Les contenus qui enfreignent les règles de Viber ou qui sont jugés autrement répréhensibles par la plateforme sont retirés. Dans ce cas, Viber peut suspendre ou résilier les comptes des utilisateurs et bloquer les participants des communautés.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.

11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. L'EIIL a créé (Site Intelligence Group Enterprise, 2018) un compte Nashir News Agency (agence de diffusion dans les médias associée à l'EIIL) sur Viber (Katz, A Growing Frontier for Terrorist Groups: Unsuspecting Chat Apps, 2019). Viber a fermé le compte immédiatement après l'avoir découvert.

27. Discord

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. Les règles de la communauté de Discord interdisent toutefois les attaques personnelles ou communautaires fondées sur des caractéristiques telles que la couleur de peau, l'origine ethnique, la nationalité, le sexe, le genre, l'orientation sexuelle, l'affiliation religieuse ou un handicap. Elles interdisent également les menaces de violence ou de préjudice à autrui. Elles défendent encore l'utilisation de Discord pour organiser, promouvoir ou soutenir l'extrémisme violent. Dans le dernier rapport de transparence de Discord, les contenus extrémistes violents sont définis comme comme étant des « contenus dans lesquels des utilisateurs font l'apologie de la violence ou y apportent leur soutien en tant que moyen à visée idéologique ». Il accompagne sa définition d'exemples tels que les groupes violents à motivation raciale, les groupes idéologiques religieux faisant usage de la violence et les groupes de célibataires involontaires se présentant comme tels, appelés « Incel ». Les conditions d'utilisation de Discord précisent en outre qu'il est interdit de diffamer, de calomnier, de tourner en ridicule, de moquer, de traquer, de harceler, d'intimider ou de maltraiter quiconque.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://discordapp.com/terms et https://discordapp.com/guidelines.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté	Discord explique qu'une infraction à la charte d'utilisation de la communauté ou à d'autres règles l'habilite à prendre « un certain nombre de mesures », définies à la section 6 ci-dessous.

(suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	
4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Les utilisateurs peuvent faire appel des décisions prises à l'encontre de leur compte.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Les utilisateurs peuvent signaler les contenus contraires à la charte d'utilisation de la communauté et aux conditions d'utilisation de Discord. Discord indique que bien qu'il ne lise pas les messages privés des utilisateurs, il mène des enquêtes et prend si nécessaire des mesures immédiates en cas de signalement d'une violation de ses conditions d'utilisation par un serveur (espace comparable à un groupe ou à une communauté, créé autour d'un thème commun) (Liao, 2018). Après le signalement, l'équipe Confiance et sécurité de Discord mène une enquête en examinant tous les éléments disponibles et en rassemblant le plus d'informations possible. L'enquête porte avant tout sur les messages signalés, mais elle peut être élargie si des éléments révèlent une infraction plus grande, par exemple si le serveur a pour seul but de se livrer à un comportement répréhensible ou que ce comportement semble s'être déjà produit par le passé. Discord utilise des « ordinateurs intelligents », des outils automatisés et des systèmes comme PhotoDNA pour détecter les messages indésirables et les contenus d'exploitation, tels que la vengeance pornographique, l'hypertrucage et les contenus menaçant la sécurité des enfants. Le dernier rapport de transparence de Discord laisse entendre que ces outils ont été utilisés pour détecter des cas d'extrémisme violent (Discord, 2020). Discord a reçu des signalements concernant des serveurs qui visaient principalement à diffuser des discours de haine, à harceler autrui et à faire l'apologie d'idéologies dangereuses. Il affirme prendre ces signalements au sérieux et retirer les serveurs affichant un comportement extrémiste (le caractère extrémiste violent n'est pas mentionné spécifiquement). Il déclare également travailler en collaboration avec les autorités répressives, des tiers (tels que des organes d'information et des universitaires) et des organisations

	spécialisées dans la lutte contre la haine (comme l'Anti-Defamation League ou le Southern Poverty Law Center) pour se tenir informé de tous les risques potentiels. Le coût économique marginal d'une utilisation des outils automatisés pour détecter les contenus discutables est probablement très faible (bien que les coûts fixes puissent être élevés), alors que celui d'un recours à des modérateurs humains pour la même fonction est probablement relativement élevé. Discord est membre du GIFCT, mais ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	Si Discord détecte une violation de sa charte d'utilisation de la communauté, il peut prendre l'une des mesures suivantes à l'encontre des utilisateurs ou des serveurs concernés : - supprimer des contenus ; - avertir l'utilisateur et lui expliquer l'infraction commise ; - exclure provisoirement l'utilisateur pour apaiser les choses ; - exclure définitivement un utilisateur et rendre difficile la possibilité de créer un nouveau compte ; - supprimer un serveur ; - désactiver la capacité d'un serveur à inviter de nouveaux utilisateurs.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Oui, mais de manière extrêmement limitée. Discord a publié son premier rapport de transparence en 2019 (Discord, 2019). Celui-ci révèle le nombre de signalements reçus à la suite d'infractions à sa charte d'utilisation de la communauté, un nombre qui pourrait inclure la publication de contenus terroristes et extrémistes violents, bien que cela ne soit pas mentionné explicitement. Le deuxième rapport de transparence de Discord (couvrant la période d'avril à décembre 2019) adopte une structure semblable et présente plusieurs informations relatives au retrait de contenus extrémistes violents (Discord, 2020).
8. Informations ou types de données figurant dans les rapports de transparence	Le deuxième rapport de transparence de Discord indique : - le nombre total de signalements reçus, ainsi que le pourcentage relevant de chaque catégorie de contenus interdits (tels que les blessures auto-infligées, le harcèlement, les menaces et les messages indésirables) ; la catégorie dont relève l'extrémisme violent n'est cependant pas clairement définie ; - le pourcentage de signalements pour lesquels Discord a pris des mesures ; il ne précise toutefois pas si celles-ci ont consisté à retirer le contenu, envoyer un avertissement ou supprimer un compte ;

	- le nombre total d'exclusions de comptes et de serveurs, par catégorie de contenus interdits ; - le nombre de suppressions de serveurs de contenus extrémistes violents par mois qui ont été détectés proactivement à l'aide d'outils automatisés ; - le nombre de comptes rétablis après recours, par catégorie de contenus interdits.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Aucune information n'est communiquée à cet égard.
10. Fréquence de publication des rapports de transparence	Cette information n'est pas définie. Discord a toutefois mentionné dans son dernier rapport de transparence qu'il a l'intention d'observer un programme de publication semestriel.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. Voir la section 8 ci-dessus.

28. Vimeo

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. Vimeo interdit toutefois les contenus qui encouragent ou soutiennent les « groupes haineux ou de terreur », qui représentent des actes illicites ou de violence extrême ou qui fournissent des instructions sur la manière de fabriquer des engins explosifs/incendiaires ou des armes artisanales/improvisées. Les membres d'un « groupe haineux ou de terreur » ne peuvent pas créer de compte Vimeo. Le terme « groupe haineux ou de terreur » n'est pas défini. Par ailleurs, un contenu est considéré comme violant la politique de lutte contre la haine et la discrimination de Vimeo lorsqu'il (1) est dirigé contre un groupe en raison de caractéristiques liées aux personnes formant ce groupe, telles que l'origine ethnique, la religion, le sexe et l'orientation sexuelle, (2) diffuse un message d'infériorité et (3) serait considéré comme extrêmement insultant par une personne raisonnable. La définition de Vimeo s'étend ainsi aux vidéos qui se prévalent de stéréotypes dangereux, affirment la supériorité raciale d'un groupe par rapport à un autre

	ou laissent entendre que certains groupes de personnes d'une religion particulière sont impliqués dans de vastes conspirations (Cheah, 2019).
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://vimeo.com/terms et https://vimeo.com/help/guidelines.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Vimeo indique que le contexte joue un rôle essentiel dans l'application de ses règles et de ses processus. Si des contenus interdits apparaissent dans le contexte d'une story ou dans la description d'une œuvre dramatique, ils seront probablement conservés, mais si l'œuvre mise en ligne vise principalement à diffuser un point de vue expressément interdit par Vimeo, elle sera retirée. Vimeo tient également compte des discours tenus par l'utilisateur concerné sur d'autres plateformes (telles que les réseaux sociaux, des blogues ou d'autres espaces où sont clairement exposées les opinions personnelles) en examinant ses intentions et sa bonne foi (Cheah, 2019). En règle générale, les modérateurs de Vimeo retirent les vidéos qui montrent des scènes de meurtres, de tortures ou d'agressions physiques ou sexuelles ou qui présentent des images choquantes, macabres ou susceptibles d'inspirer du dégoût. Vimeo comprend néanmoins que des vidéos traitent de ces sujets d'une manière réfléchie et critique. Les vidéos qui filment le monde réel comprennent parfois des scènes violentes. Le contexte est donc essentiel et les vidéos documentaires ou journalistiques disposent d'une plus grande latitude pour montrer des scènes de violence ou leurs répercussions. Pour ne pas être supprimées, ces vidéos ne doivent pas comporter d'images à sensation, gratuites ou d'exploitation. Elles doivent aussi porter la mention « adulte ». Les vidéos qui recrutent pour des organisations terroristes ou font leur propagande, qu'elles montrent

	ou non des scènes de violence, ne sont jamais autorisées (Vimeo, n.d.).
4.1 Notifications des suppressions ou des autres décisions de sanction	Certaines décisions de retrait de contenu font l'objet d'une notification, telles que celles concernant les violations de droits d'auteur. Vimeo n'informe toutefois pas les utilisateurs concernés du retrait d'une vidéo ou d'un compte (et ne propose pas de mécanisme de recours) si les contenus retirés relèvent de certaines catégories, telles que la suspicion de maltraitance d'enfants ou le terrorisme.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Il est possible de faire appel des retraits pour violation de droits d'auteur, mais il n'existe pas de mécanisme de recours en cas de retrait de contenus terroristes et extrémistes violents.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Les utilisateurs peuvent signaler les contenus contraires aux lignes de conduite et aux règles de Vimeo. Vimeo indique qu'il peut surveiller les comptes, les contenus et le comportement des utilisateurs, quels que soient leurs paramètres de confidentialité. Vimeo a conclu un accord avec Active Fence pour participer à la détection des contenus terroristes et extrémistes violents, et devrait mettre en œuvre ce partenariat début 2020. Le coût économique marginal d'un recours à des modérateurs humains pour détecter les contenus répréhensibles est probablement relativement élevé. Vimeo n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	En cas de violation de ses règles et conditions d'utilisation, Vimeo peut, à sa seule discrétion, résilier, suspendre ou limiter l'accès de l'utilisateur concerné à son compte ou à ses contenus, et clôturer le compte concerné.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.

9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Information inconnue.

29. IMO

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. La politique d'utilisation acceptable d'IMO interdit toutefois l'utilisation de ses services pour diffuser des menaces de violence.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://imo.im/policies/terms_of_service et https://imo.im/policies/acceptable_use_policy.html.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	IMO indique se réserver le droit de supprimer, de filtrer, de modifier ou de désactiver l'accès à des contenus sans en avertir le propriétaire au préalable s'il considère, à sa seule discrétion, qu'ils enfreignent ses règles ou nuisent de quelque manière que ce soit à ses services.

4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Aucune procédure de recours n'est indiquée.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	IMO indique qu'il n'est soumis à aucune obligation d'examiner les contenus, mais qu'il se réserve le droit de le faire à tout moment. Il ne précise toutefois pas quels types d'examens sont réalisés. IMO n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	La violation des règles d'IMO peut entraîner la suspension ou la clôture du compte de l'utilisateur concerné.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.

11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Information inconnue.

30. LINE

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. Les conditions d'utilisation de LINE interdisent toutefois la publication ou la diffusion de contenus violents. Les activités qui bénéficient à des groupes antisociaux ou qui sont menées dans le cadre d'une collaboration avec ces derniers sont également interdites. Le terme « groupe antisocial » n'est pas défini.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://terms.line.me/line_terms/.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Oui. Les textes sont disponibles sur https://terms2.line.me/LINELIVE_ToC_ME1.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	LINE applique un processus composé de deux étapes pour surveiller les publications mises en ligne sur son fil d'actualité et sur LINE LIVE, LINE Manga, LINE Fortune, LINE Pasha, LINE Step, LINE BLOG, LINE Delima et WizBall. Le contenu publié par un utilisateur sur l'un des services LINE est tout d'abord contrôlé par le système de surveillance automatique de la plateforme pour vérifier qu'il ne comprend pas de mots interdits, qu'il n'enfreint aucune règle de la plateforme et respecte ses conditions d'utilisation ainsi que la législation en vigueur. Si un contenu répréhensible est détecté par le système de surveillance, il est suspendu dès sa mise en ligne. Une équipe de surveillance vérifie ensuite les contenus que le système automatique ne peut pas traiter. Elle les évalue en fonction d'une série de critères et les compare à des exemples de contenus rencontrés précédemment pour déterminer s'ils sont autorisés ou non. Si elle décide qu'ils sont contraires aux conditions d'utilisation de LINE ou à la législation en vigueur, ils sont suspendus (LINE, 2019-2020).

	LINE ne peut pas surveiller les messages envoyés ou reçus par un utilisateur dans un salon de discussion LINE classique, sauf si l'utilisateur lui communique les données non chiffrées de la conversation en utilisant l'outil de signalement (LINE, 2019-2020) .
4.1 Notifications des suppressions ou des autres décisions de sanction	Il n'existe pas de notification de retrait de contenu.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Les utilisateurs peuvent faire appel des décisions de retrait par l'intermédiaire du formulaire de contact de LINE.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Les utilisateurs peuvent signaler les contenus contraires aux règles de LINE. Les signalements sont examinés par l'équipe de LINE, qui « prend des mesures appropriées » (LINE, n.d.) si elle constate une violation de ces règles. En plus de traiter les signalements effectués par les utilisateurs, l'équipe et le système automatique de surveillance de LINE examinent avec sérieux les publications des utilisateurs (voir la section 4 ci-dessus). Le coût économique marginal d'une utilisation des outils automatisés pour détecter les contenus terroristes et extrémistes violents est probablement très faible (bien que les coûts fixes puissent être élevés), alors que celui d'un recours à des modérateurs humains pour la même fonction est probablement relativement élevé. LINE n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	LINE peut supprimer les contenus ou suspendre ou supprimer un compte sans avertissement préalable s'il estime que l'utilisateur enfreint ou a enfreint ses règles.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non. LINE publie néanmoins des rapports de transparence sur trois sujets : les demandes de communication des données utilisateurs ou de suppression émanant des autorités répressives, les mesures prises à l'encontre des publications qui enfreignent ses conditions d'utilisation ou la législation en vigueur et l'état de déploiement du chiffrement des messages et des appels (LINE, 2019-2020).

8. Informations ou types de données figurant dans les rapports de transparence	Le rapport sur les mesures prises à l'encontre des publications en infraction sur les services LINE indique le nombre de contenus suspendus et leur pourcentage par catégorie, à savoir les messages indésirables, les contenus obscènes, les sollicitations, l'utilisation commerciale non autorisée des comptes, les contenus gênants et problématiques, la promotion d'activités illégales, ainsi qu'une catégorie « autres ». Les contenus terroristes et extrémistes violents semblent relever de la catégorie « promotion des activités illégales » (d'après les exemples cités à la section 9 ci-dessous), mais cela n'est pas indiqué de manière explicite.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	LINE précise que les contenus gênants et problématiques peuvent être « des remarques excessivement haineuses, des photos de cadavres, de la fraude au clic, des liens vers des sites de hameçonnage, etc. » et que la promotion les activités illégales peut recouvrir « des annonces d'attaques ou d'attentats, la vente de stupéfiants, la vente de données en ligne (tels que des comptes, des monnaies et des avatars) contre de l'argent, etc. ».
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Information inconnue.

31. Huoshan

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition particulière des contenus terroristes et extrémistes violents. Les conditions d'utilisation de Huoshan interdisent toutefois la promotion du terrorisme et de l'extrémisme (sans mentionner spécifiquement l'extrémisme violent).
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.huoshanzhibo.com/agreement/.
3. Présence de dispositions précises applicables aux contenus	Non.

diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Aucune procédure n'est indiquée. Huoshan indique qu'il tient un registre des infractions présumées à la législation et aux réglementations, qu'il signale ces cas aux autorités compétentes conformément à la loi et qu'il coopère à toutes les enquêtes relatives à ces cas.
4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Aucune procédure de recours n'est indiquée.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Les utilisateurs peuvent signaler tous types d'activités ou de contenus illicites sur Huoshan. L'équipe de modérateurs de Huoshan examine les signalements et, le cas échéant, prend des mesures adaptées. Huoshan possède aussi une équipe spécialisée dans la modération de contenus et accentue ses efforts pour améliorer ses « normes de contrôle » (Yoo, 2018). Le coût économique marginal d'un recours à des modérateurs humains pour détecter les contenus répréhensibles est probablement relativement élevé. Huoshan n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions	En cas de violation de ses conditions d'utilisation, Huoshan peut supprimer les publications ou les commentaires concernés, restreindre tout ou partie des fonctions du compte concerné ou résilier l'accès à ses services.

d'utilisation ou des règles de la communauté	
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Information inconnue.

32. Ask.fm

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. Les règles de la communauté d'Ask.fm indiquent toutefois que les organisations terroristes et les groupes extrémistes violents qui cherchent à encourager ou à commettre des activités terroristes ou criminelles violentes ne sont pas autorisés à être présents sur la plateforme pour promouvoir leurs campagnes ou leurs plans, célébrer leurs actes de violence, lever des fonds ou recruter des jeunes. Les termes « organisations terroristes » et « groupes extrémistes violents » ne sont pas définis. Par ailleurs, les utilisateurs ne peuvent pas publier de contenus comportant des menaces de quelque nature qu'elles soient, telles que des menaces de violence physique envers eux-mêmes ou autrui, ou incitant à commettre des actes de violence envers eux-mêmes ou autrui.

	Les mots « terroriste », « terrorisme » ou « extrémisme » ne sont pas explicitement définis.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://about.ask.fm/legal/2019-07/en/terms.html et https://about.ask.fm/community-guidelines/.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Sans objet. Ask.fm ne propose pas de fonctionnalité de diffusion en direct de contenus.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Ask.fm indique se réserver le droit de surveiller l'accès des utilisateurs à ses services ou leur utilisation pour vérifier qu'ils respectent ses conditions d'utilisation, et d'examiner ou de modifier les contenus. Il stipule également qu'il peut bloquer ou désactiver l'accès à des contenus considérés comme répréhensibles ou nuisibles pour autrui sans avertissement préalable.
4.1 Notifications des suppressions ou des autres décisions de sanction	Les contenus qui ne respectent pas les conditions d'utilisation ou les règles de la communauté d'Ask.fm sont retirés, suite à quoi leur auteur reçoit un avertissement écrit. La suspension ou la résiliation prochaines de l'accès aux services ou au profil de l'utilisateur d'Ask.fm est notifié à l'utilisateur dans des conditions raisonnables. L'avertissement peut être envoyé plusieurs fois automatiquement par le système ou manuellement par un modérateur avant que le blocage du profil intervienne.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Les utilisateurs dont le compte a été supprimé peuvent faire appel de cette décision.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage	Les utilisateurs peuvent signaler les contenus qu'ils estiment contraires aux règles d'Ask.fm. Les signalements sont transmis à l'équipe d'Ask.fm pour examen. Ask.fm affirme examiner tous les signalements effectués et indique qu'il peut consulter les informations et les contenus d'un utilisateur s'il estime raisonnablement nécessaire de mettre en œuvre ses conditions d'utilisation et de protéger la sécurité de ses utilisateurs ou du public.

d'empreintes numériques ou d'adresses URL)	Le coût économique marginal d'un recours à des modérateurs humains pour détecter les contenus répréhensibles est probablement relativement élevé. Ask.fm a entamé les démarches pour devenir membre du GIFCT.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	La violation des conditions d'utilisation d'Ask.fm peut entraîner la suspension ou la clôture du compte de l'utilisateur concerné ou la résiliation de son accès aux services.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. Un compte Ask.fm aurait par exemple fourni des conseils sur la manière de rejoindre les combattants de l'EIIL en Iraq, ainsi que des informations sur les armes fournies à l'arrivée. (Miller, 2014)

33. YY Live

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition particulière des contenus terroristes et extrémistes violents. Les conditions d'utilisation de YY Live stipulent toutefois que les utilisateurs ne peuvent pas publier, transmettre, diffuser et conserver des contenus violents ou des contenus faisant la promotion du terrorisme, de l'extrémisme et des activités connexes.

2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://zc.yy.com/license.html.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Aucune procédure n'est indiquée.
4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Aucune procédure de recours n'est indiquée.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Aucune information n'est communiquée. Des recherches ont toutefois montré que YY Live applique une surveillance et une censure par mot clé. (Knockell, 2015) Pour faire respecter ses conditions d'utilisation, YY Live possède dans son service de sécurité des données une équipe qui assure une surveillance 24 heures sur 24 des contenus en s'appuyant sur un système qui balaie la plateforme à la recherche de contenus inappropriés et procède à un filtrage automatique sur la base de mots clés. (Knockell, 2015) Le coût économique marginal d'une utilisation des outils automatisés pour détecter les contenus discutables est probablement très faible (bien que les coûts fixes puissent être élevés), alors que celui d'un recours à des modérateurs humains pour la même fonction est probablement relativement élevé.

	YY Live n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	En cas de violation de ses conditions d'utilisation, YY Live peut limiter ou bloquer l'accès de l'utilisateur concerné à son compte et restreindre ou suspendre son accès à un ou plusieurs produits, services ou fonctions (les vidéos en direct, par exemple).
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Information inconnue.

34. Twitch

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Twitch a mis à jour ses lignes de conduite « Terrorisme et violence extrême » en octobre 2020 pour indiquer plus spécifiquement qu'il interdit les contenus terroristes et extrémistes violents. Twitch n'autorise pas les contenus qui représentent, font l'apologie, encouragent ou soutiennent le terrorisme, ou les individus ou actes extrémistes violents. Cela inclut le fait de menacer ou d'encourager d'autres personnes à commettre des actes qui entraîneraient des dommages physiques graves chez des groupes de personnes ou une destruction importante de biens. Twitch étend l'interdiction à l'affichage ou à la publication de liens de propagande terroriste ou extrémiste, y compris des images ou des séquences graphiques de violence terroriste ou extrémiste, et ce même dans le but de dénoncer un tel contenu. (Les lignes de conduite de Twitch sont

	disponibles à l'adresse https://www.twitch.tv/p/en/legal/community-guidelines/.) Ces clarifications de la mise à jour ont élargi le champ de la catégorie « Terrorisme et violence extrême » (en y englobant des formes de comportement qui relevaient jusque-là des autres types d'abus), ce qui a entraîné une augmentation substantielle des sanctions de ce type en pourcentage, mais pas nécessairement en chiffre absolu (voir à ce sujet la section 9 ci-dessous). Les lignes de conduite de la communauté de Twitch stipulent également que les actes et menaces de violence seront pris au sérieux et sont considérés comme des violations qui seront traitées avec une tolérance zéro. Tous les comptes associés à de telles activités seront suspendus pour une durée indéterminée. Cela inclut notamment : • les tentatives ou menaces d'attaque physique ou de meurtre à l'encontre d'autres personnes ; • l'utilisation d'armes pour menacer, intimider, blesser physiquement ou tuer d'autres personnes. Twitch interdit encore les comportement haineux, ceux-ci désignant tout contenu ou toute activité qui promeut ou encourage la discrimination, le dénigrement, le harcèlement ou la violence sur la base des caractéristiques protégées suivantes : race, ethnie, couleur, caste, nationalité, statut d'immigration, religion, sexe, genre, identité sexuelle, orientation sexuelle, handicap, état de santé grave et statut d'ancien combattant. Il offre également certaines protections pour l'âge. Twitch applique une tolérance zéro en ce qui concerne les comportements haineux, indiquant par là qu'il prend des mesures pour chaque cas de comportement haineux valable signalé. En vertu de cette politique, il offre à tous les utilisateurs les mêmes protections, quelles que soient leurs caractéristiques individuelles.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.twitch.tv/p/en/legal/community-guidelines/, https://www.twitch.tv/p/en/legal/terms-of-service/ et https://help.twitch.tv/s/topic/0TO1U000000CjnZWAS/moderation-safety?language=en_US.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Comme Twitch est avant tout un service de diffusion en direct de contenus, ses conditions et règles sont adaptées spécialement aux contenus diffusés en direct et s'y appliquent directement.

4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Twitch prend des sanctions à l'encontre des comptes qui enfreignent ses conditions d'utilisation ou ses lignes de conduite de la communauté. Il prend en compte plusieurs facteurs lors de l'examen des signalements d'infraction, notamment l'intention et le contexte, les dommages potentiels occasionnés à la communauté, les obligations légales et d'autres éléments. Selon la nature de l'infraction, Twitch applique différentes sanctions qui comprennent l'envoi d'un avertissement, la suspension provisoire d'un compte et l'exclusion définitive pour les infractions les plus graves. Un avertissement est un préavis. Twitch peut aussi retirer les contenus associés à l'infraction. La répétition d'une infraction pour laquelle l'utilisateur a déjà reçu un avertissement, ou la commission d'une infraction semblable, entraînera une suspension. Les suspensions temporaires durent de 24 heures à des périodes pouvant dépasser 30 jours. En cas de suspension, l'utilisateur ne peut plus consulter ni utiliser les services de Twitch, notamment regarder ou diffuser des vidéos en direct, converser en ligne, créer d'autres comptes et apparaître ou participer sur la chaîne d'un tiers. Il peut à nouveau utiliser les services de Twitch à l'issue de la suspension. Twitch archive les infractions, et plusieurs suspensions consécutives peuvent entraîner une exclusion définitive. Pour les infractions les plus graves, Twitch suspend le compte immédiatement et définitivement.
4.1 Notifications des suppressions ou des autres décisions de sanction	Les décisions de sanction sont communiquées par courrier électronique à la personne qui a mis le contenu en ligne. Les notifications précisent le type de contenu retiré, indiquent où ce retrait a eu lieu et fournissent des exemples du comportement incriminé. Elles renvoient en outre aux lignes de conduite de la communauté (à la section concernée) et précisent la durée des sanctions imposées à la suite de l'infraction. Elles indiquent également comment faire appel des décisions de sanction. Il est à noter que Twitch n'avertit pas l'auteur de la mise en ligne du contenu en cas d'activité illégale ou lorsqu'une notification pourrait compromettre l'enquête éventuelle des autorités compétentes.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Si l'utilisateur pense qu'il n'a pas enfreint les lignes de conduite de la communauté de Twitch, il peut faire appel de la décision de sanction. Il doit indiquer dans son recours la raison pour laquelle il estime que la décision n'est pas justifiée. Lorsque l'appel a été examiné, Twitch informe l'utilisateur de son issue.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu	Twitch met à la disposition des utilisateurs des outils leur permettant de signaler des contenus ou des comportements qui enfreignent ses lignes de conduite de la communauté, que ce soit dans les vidéos diffusées en direct, dans le dialogue en ligne ou dans des commentaires associés à une vidéo. Twitch a également mis en place à l'échelle du service des technologies de « détection automatique » qui repèrent les contenus liés à la nudité, ceux à caractère sexuel ou à caractère sanglant et les contenus

généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	correspondant à de la violence extrême pour les soumettre à une évaluation. Twitch bloque les noms d'utilisateur qui enfreignent ses règles de la communauté. Les signalements sont examinés par l'équipe de sécurité de Twitch, ceux faisant état de violence extrême ou relevant du terrorisme étant traités en priorité. Un deuxième niveau de modération repose sur une série d'outils qui permettent au propriétaire d'une chaîne (parfois appelé « diffuseur ») de nommer d'autres utilisateurs modérateurs de la chaîne. Ces derniers sont alors en mesure d'exclure les utilisateurs irrespectueux, de bloquer certains mots ou expressions, d'exiger une vérification par téléphone et de retirer des messages du dialogue en ligne. Ils peuvent prendre les mêmes mesures que le propriétaire de la chaîne. Enfin, Twitch propose aux propriétaires de chaînes un outil utilisant des algorithmes d'apprentissage automatique et de traitement du langage naturel pour empêcher l'affichage des messages pour les autres utilisateurs du dialogue en ligne tant qu'ils n'ont pas été examinés par un modérateur de la chaîne. Ce mécanisme est appelé « AutoMod » (Twitch, n.d.). AutoMod se concentre sur différentes catégories de contenus, à savoir ceux relevant de la discrimination, les contenus à caractère sexuel et les messages à caractère hostile ou vulgaires. Le coût économique marginal d'une utilisation des outils automatisés pour détecter les contenus discutables est probablement très faible (bien que les coûts fixes puissent être élevés), alors que celui d'un recours à des modérateurs humains pour la même fonction est probablement relativement élevé. Les utilisateurs modérateurs n'engendrent aucun coût pour Twitch. Twitch appartient à Amazon, qui a rejoint le GIFCT en septembre 2019.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	La violation des lignes de conduite de la communauté de Twitch peut entraîner le retrait des contenus, l'envoi d'un avertissement ou la suspension du compte. Les infractions graves sont sanctionnées par une suspension immédiate du compte.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Oui. Twitch a publié son premier rapport de transparence en février 2021, celui-ci couvrant l'intégralité de l'année 2020 (Twitch, 2020). La section « Signalements et sanctions » de son rapport de transparence présente les informations relatives aux contenus terroristes et extrémistes violents. Twitch rapporte qu'il n'a connu aucun cas d'activité terroriste diffusée en direct en 2020.
8. Informations ou types de données	Les informations présentées pour les premier et second semestres de la période considérée sont les suivantes :

figurant dans les rapports de transparence	- pourcentage de couverture de la modération dans les chaînes par l'outil AutoMod, les modérateurs des chaînes et les deux ; - nombre de retraits manuels et « proactifs » (c'est-à-dire à l'aide d'outils automatiques comme Termes bloqués et AutoMod) de messages de conversation par les modérateurs des chaînes ; - nombre de bannissements permanents et de suspensions temporaires de chaînes imposés par les modérateurs des chaînes ; - nombre consolidé de signalements d'utilisateurs à la suite d'infractions relevant des catégories suivantes ; terrorisme, propagande terroriste et endoctrinement, nudité des adultes, pornographie et comportement sexuel, violence, contenus à caractère sanglant, menaces ou autre contenu choquant, comportement haineux, harcèlement sexuel et harcèlement, recours à des robots de visionnage, envoi de messages indésirables et autres infractions aux lignes de conduite de la communauté ; - nombre total de sanctions ; - nombre de sanctions relevant de la catégorie Comportement haineux, harcèlement sexuel et harcèlement ; - nombre de sanctions relevant de la catégorie Violence, contenus à caractère sanglant, menaces ou autre contenu choquant ; - nombre de sanctions relevant de la catégorie Nudité des adultes, pornographie et comportement sexuel ; - nombre de sanctions relevant de la catégorie Messages indésirables et autres infractions aux lignes de conduite de la communauté ; - nombre de sanctions relevant de la catégorie Terrorisme, propagande terroriste et endoctrinement, subdivisée en sanctions pour la diffusion de propagande terroriste et la glorification ou l'encouragement d'actes de terrorisme, de violence extrême ou de destruction de biens à grande échelle.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Twitch explique que la grande majorité des suppressions de contenu sur Twitch concerne des suppressions de messages de chat effectuées par des modérateurs sur des chaînes individuelles. Il convient toutefois de rappeler que Twitch est un service de diffusion de contenus en direct et que la grande majorité du contenu y est éphémère. C'est pourquoi Twitch ne se concentre pas sur le « retrait de contenu » comme principal moyen pour faire respecter ses lignes de conduite de la communauté par les diffuseurs. Le contenu en direct est en fait signalé par la détection automatique ou par des utilisateurs à l'équipe de professionnels de la modération de contenu de Twitch, qui imposent alors des « sanctions » (par exemple, un avertissement ou une suspension temporaire de la chaîne) pour remédier aux infractions avérées. Si une violation s'accompagne de contenu enregistré, ce contenu est également retiré. Il est à noter cependant que la plupart des sanctions n'entraînent pas la suppression du contenu : en effet, en dehors du signalement, il n'y a plus de traces de l'infraction puisque le contenu en direct a déjà disparu. Twitch considère donc que la mesure la plus appropriée

	qu'elle puisse prendre dans le cadre de ses efforts pour la sécurité consiste en des « sanctions », d'où la prépondérance de cet indicateur dans son rapport de transparence. Par souci de clarté, Twitch fait observer que les statistiques relatives aux sanctions de la section « Signalements et sanctions » de son rapport de transparence ne comprennent pas les sanctions prises au niveau des chaînes traitées à la section « Modération sur les chaînes : couverture du service et mécanismes de suppression et de sanction », et qu'elles n'en constituent pas des doublons.
10. Fréquence de publication des rapports de transparence	La fréquence de publication est annuelle.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. Lors d'une attaque coordonnée sur Twitch en mai 2019, des utilisateurs ont diffusé des contenus choquants, notamment des extraits de la vidéo de l'attentat de Christchurch. (Marshall, 2019) En octobre 2019 aussi, l'auteur de l'attentat de Halle, en Allemagne, a diffusé son attaque en direct sur Twitch. (British Broadcasting Corporation (BBC), 2019) Celle-ci a été vue par environ 2 500 personnes avant que Twitch ne retire la vidéo, qui n'a ensuite pas réapparu.

35. Tumblr

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition particulière des contenus terroristes et extrémistes violents. Les règles communautaires de Tumblr indiquent toutefois que les contenus qui promeuvent, incitent ou encouragent les actes terroristes ne sont pas tolérés. Sont compris les contenus qui soutiennent ou glorifient des organisations terroristes, leurs dirigeants ou les activités violentes qui y sont liées. Le terme « organisation terroriste » n'est pas défini. Tumblr interdit également les propos haineux, définis comme étant un contenu promouvant la haine ou incitant à la haine contre des personnes ou des groupes en fonction de caractéristiques telles que la race, l'origine ethnique ou la nationalité, la religion, le sexe, l'identité de genre, l'âge, le statut de vétéran, l'orientation sexuelle, le handicap ou la maladie, ou déshumanisant ces personnes ou ces groupes au titre de ces caractéristiques.
2. Manière dont les conditions d'utilisation ou les règles de la	Les textes sont disponibles sur https://www.tumblr.com/policy/en/terms-of-service et https://www.tumblr.com/policy/en/community.

communauté sont communiquées	
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Si Tumblr conclut qu'un utilisateur a enfreint ses règles, il lui envoie un avertissement par courrier électronique. Si l'utilisateur refuse de se justifier ou de corriger son comportement, Tumblr peut prendre des mesures contre son compte. La plateforme précise qu'elle se réserve le droit de suspendre un compte ou de supprimer du contenu, sans préavis, quelle qu'en soit la raison, pour protéger ses services, ses infrastructures, ses utilisateurs ou sa communauté.
4.1 Notifications des suppressions ou des autres décisions de sanction	Il n'existe pas de notification de retrait de contenu.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Les utilisateurs peuvent contacter le centre d'aide de Tumblr pour faire appel d'une décision de retrait de contenu.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Les utilisateurs peuvent signaler les activités ou les contenus illégaux sur Tumblr. Tumblr indique que ses spécialistes formés examinent les contenus signalés et prennent des « mesures appropriées ». Les signalements n'entraînent pas systématiquement le retrait des contenus concernés, les équipes spécialisées de Tumblr estimant parfois que le contenu signalé n'enfreint pas ses règles communautaires. Outre les signalements envoyés par les utilisateurs, Tumblr utilise des outils automatisés pour détecter les contenus pouvant être associés à du terrorisme ou de l'extrémisme violent et les soumettre à un examen humain.

	Le coût économique marginal d'une utilisation des outils automatisés pour détecter les contenus répréhensibles est probablement très faible (bien que les coûts fixes puissent être élevés), alors que celui d'un recours à des modérateurs humains est probablement relativement élevé. Tumblr n'est pas membre du GIFCT, mais participe au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.[51]
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	Tumblr peut résilier ou suspendre l'accès de l'auteur de l'infraction aux services ou sa capacité à les utiliser, immédiatement et sans avertissement préalable, et décline toute responsabilité.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non. Oath, qui détenait auparavant Tumblr (Alexander, 2019), publie des rapports de transparence qui, jusqu'en 2018, portaient également sur le service Tumblr. Ils sont toutefois très larges et ne répartissent pas les informations entre les différentes sociétés contrôlées par Oath (les demandes de retrait des autorités concernent par exemple à la fois Yahoo et Tumblr). Ils ne comportent pas non plus d'informations spécifiques sur les contenus terroristes et extrémistes violents. (Verizon Media, 2019). En 2019, Tumblr a été vendu à Automattic. Depuis, un rapport de transparence (se rapportant au deuxième semestre de 2019) a été publié, mais il porte uniquement sur les demandes de données d'utilisateur et de retraits de contenus émanant des autorités publiques. Il ne comporte pas d'informations sur les contenus terroristes et extrémistes violents. (Tumblr, 2019)
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus	Oui. Tumblr comporterait un grand nombre de pages faisant l'apologie du nazisme, du suprémacisme blanc, du nationalisme ethnique et du terrorisme d'extrême droite (Barnes, 2019) (Fisher-Birch, 2018).

terroristes et extrémistes violents	

36. Flickr

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition particulière des contenus terroristes et extrémistes violents. Les règles de la communauté de Flickr interdisent toutefois la publication de contenus associés au terrorisme. Flickr applique également une politique de tolérance zéro pour les attaques visant des personnes ou des groupes sur la base, notamment, de leur origine ou appartenance ethnique, nationalité, religion, handicap, pathologie, âge, orientation sexuelle, sexe ou identité de genre.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.flickr.com/help/terms et https://www.flickr.com/help/guidelines.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Flickr s'appuie sur des utilisateurs modérateurs pour détecter les contenus comprenant des éléments de nudité ou indécents, mais ce système ne s'applique pas aux contenus terroristes et extrémistes violents dans la mesure où la publication d'un contenu de cette nature entraîne la suppression du compte concerné. Les critères de détection des contenus terroristes et extrémistes violents ne sont toutefois pas précisés.
4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.
4.2 Mécanismes de recours en cas de suppression ou	Aucune procédure de recours n'est indiquée.

d'autres décisions de sanction	
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Les utilisateurs peuvent signaler les contenus qu'ils estiment contraires aux règles de la communauté de Flickr. Le personnel de Flickr examine les signalements pour déterminer si le contenu enfreint effectivement les règles et, le cas échéant, prend des mesures. Le coût économique marginal d'un recours à des modérateurs humains pour détecter les contenus répréhensibles est probablement relativement élevé. Flickr n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	La publication de contenus terroristes et extrémistes violents entraîne la suppression du compte concerné. Flickr indique qu'il peut signaler ces comportements aux autorités répressives.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. Un monument virtuel a été créé sur la plateforme pour les combattants du *djihad* tués en Syrie, avec leur nom et leur origine, ainsi que des commentaires louant leur dévotion et leur force de combat (Weimann, 2014).

37. VK

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition particulière des contenus terroristes et extrémistes violents. Les conditions d'utilisation de VK interdisent toutefois aux utilisateurs de mettre en ligne, conserver, publier, diffuser, mettre à disposition ou utiliser de quelque manière que ce soit des informations qui contiennent des éléments extrémistes et qui encouragent des activités criminelles, ou qui fournissent des conseils, des instructions ou des guides relatifs à des activités criminelles. Les règles de la plateforme VK interdisent aux utilisateurs de publier des contenus faisant la promotion d'activités illicites, d'organisations criminelles ou du terrorisme. Elles interdisent également les contenus incitant à la haine ou à l'hostilité raciale, religieuse ou ethnique ou les propageant, y compris la haine ou l'hostilité envers les personnes d'un sexe, d'une orientation ou d'autres attributs ou caractéristiques particuliers (y compris en matière de santé). VK reprend la définition de « contenus terroristes ou extrémistes violents » de la législation des pays dans lesquels il est présent.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://vk.com/terms, https://vk.com/licence et https://m.vk.com/safety?lang=en&section=standarts.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Les conditions émanant des conditions d'utilisation (https://vk.com/terms), du contrat de licence (vk.com/licence) et des règles de la communauté (vk.com/safety?section=standards) de VK s'appliquent tout autant aux contenus diffusés en direct qu'aux autres types de contenus.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de	Aucune procédure particulière n'est indiquée. VK indique de manière générale qu'il se réserve le droit, à sa seule discrétion ainsi qu'à la suite de la réception d'informations de la part d'utilisateurs ou de tiers, de modifier (modérer), bloquer ou retirer tout contenu publié qui enfreint ses conditions d'utilisation, ou de suspendre, limiter ou résilier l'accès de l'auteur de l'infraction à tout ou partie de ses services à tout moment, avec ou sans préavis. VK se réserve également le droit de supprimer la page d'un utilisateur ou de suspendre, de limiter ou de résilier l'accès de l'utilisateur à ses services s'il pense que l'utilisateur représente une menace pour la plateforme ou pour les utilisateurs.

recours contre ces décisions	
4.1 Notifications des suppressions ou des autres décisions de sanction	Les suppressions de contenus sont notifiées aux utilisateurs, y compris celles concernant les contenus figurant sur la liste fédérale des contenus extrémistes du ministère de la Justice de la Fédération de Russie.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Si un utilisateur conteste la suppression ou le blocage d'un contenu, il peut contacter le centre d'aide de VK.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	VK utilise une méthode de modération hybride. Il répond aux signalements envoyés par les utilisateurs, les agences de régulation et les autres organismes, et effectue également une surveillance en interne avec des « mécanismes de recherche automatique et de suppression des contenus inappropriés ». Parmi les différents outils automatisés à sa disposition, il utilise par exemple les empreintes numériques pour localiser rapidement les contenus nuisibles ainsi que les réseaux neuronaux. VK note que la majorité des « contenus dangereux » sont supprimés avant toute consultation (VK, 2020). Toute personne peut signaler un contenu illégal, insultant ou trompeur en cliquant sur un bouton de signalement (*Report*). L'équipe de modération intervient le plus rapidement possible pour exclure les auteurs d'infraction et bloquer les contenus contraires aux règles de la plateforme ou à la législation en vigueur. VK permet aussi aux utilisateurs de créer des « communautés » et d'en devenir administrateurs et modérateurs. D'après les conditions d'utilisation de VK, il est de la responsabilité des administrateurs et des modérateurs d'une communauté de modérer et de bloquer les contenus mis en ligne sur les pages contrôlées par leur communauté. Ils doivent en particulier supprimer tout contenu contraire aux conditions d'utilisation ou à la législation en vigueur. Le coût économique marginal d'une utilisation des outils automatisés pour détecter les contenus discutables est probablement très faible (bien que les coûts fixes puissent être élevés), alors que celui d'un recours à des modérateurs humains pour la même fonction est probablement relativement élevé. VK n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation	En cas de violation des conditions d'utilisation, y compris lors de la création et de l'administration d'une communauté, VK peut retirer ou supprimer les contenus en infraction, bloquer provisoirement l'accès de l'utilisateur concerné aux services,

ou des règles de la communauté	exclure le contenu des résultats des recherches ou résilier le compte concerné.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non. Toutefois, dans son centre de sécurité et ses règles de la plateforme, VK fait état de plusieurs chiffres concernant les infractions à ses règles, mais aucune information n'est donnée sur les contenus terroristes et extrémistes violents. Voir (VK, 2020) et. (VK, 2020)
8. Informations ou types de données figurant dans les rapports de transparence	Nombre d'éléments de contenu, de profils et de communautés bloqués à la suite de faits relevant de la promotion du suicide, de la violence à l'école (statistiques de 2019), de la haine ou des messages à caractère hostile (statistiques relatives aux premier et deuxième trimestres de 2020) et du trafic de stupéfiants (statistiques de 2019 ; https://vk.com/safety?section=health).
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Aucune information n'est communiquée.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. Des comptes de l'EILL ont été découverts sur VK (Lokot, 2014).

38. Medium

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. Les conditions d'utilisation de Medium stipulent toutefois que la plateforme interdit les contenus ou actions qui menacent, encouragent ou prônent la violence de manière directe ou indirecte, les contenus qui promeuvent la violence ou la haine sur la base de critères tels que la race, l'ethnie, la nationalité, la religion, le handicap, la maladie, l'âge, l'orientation sexuelle, le sexe ou l'identité sexuelle, les publications ou les comptes qui glorifient, célèbrent, minimisent ou banalisent la violence, la souffrance, les agressions ou la mort de personnes ou de groupes, ainsi que les appels à l'intolérance, l'exclusion ou la ségrégation sur la base de caractéristiques faisant l'objet d'une protection.

	L'apologie des groupes se livrant à ces comportements est également interdite.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur : https://medium.com/policy/medium-rules-30e5502c4eb4 et sur :https://medium.com/policy/medium-terms-of-service-9db0094a1e0f
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Pour les contenus signalés par les utilisateurs, Medium prend en compte des facteurs tels que l'intérêt, le contexte et la nature des informations publiées, la probabilité raisonnable, l'ampleur et la gravité du préjudice social prévisible et la législation en vigueur. Lors de l'évaluation des contenus controversés et extrémistes (le caractère extrémiste violent n'est pas mentionné spécifiquement) au regard des règles, les modérateurs employés par Medium appliquent une analyse des risques qui répond au minimum aux questions suivantes : - Quelles sont les conséquences négatives prévisibles de la diffusion de l'information par Medium et de son partage sur d'autres réseaux sociaux ? - Quelle pourrait être la gravité de ses répercussions ? - Quelle est la probabilité que les conséquences négatives se produisent ? - Qui toucheront-elles probablement ? - Des informations provenant d'institutions nationales ou internationales reconnues (telles que l'ECDC, l'OMS et d'autres organes officiels) peuvent-elles nous aider à déterminer si le contenu présente un risque élevé ? (Medium, n.d.) Medium fournit les exemples suivants de types de contenus présentant un risque élevé, plus susceptibles d'entraîner une suspension de compte ou une restriction de diffusion : - Allégations pseudo-scientifiques affirmant la supériorité ou l'infériorité d'un groupe particulier (sur la base de critères tels que la race, l'origine ethnique ou le sexe). - Théories du complot ayant déjà donné lieu à des incidents de harcèlement ou de violence entre les utilisateurs, théories dont on peut prévoir qu'elles peuvent encourager ou provoquer le harcèlement ou des préjudices physiques ou de réputation. (Medium, n.d.)

4.1 Notifications des suppressions ou des autres décisions de sanction	Medium avertit l'utilisateur lorsqu'une enquête est menée sur un contenu de son compte ou que celui-ci est désactivé, sauf s'il pense qu'il s'agit d'un compte robotisé ou opérant de mauvaise foi, ou que la notification pourrait causer un préjudice à quelqu'un, l'entretenir ou l'exacerber.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Si l'utilisateur pense que son contenu ou son compte a été limité ou désactivé par erreur ou que Medium n'a pas eu connaissance d'un élément de contexte particulier lors de sa prise de décision, il peut déposer un recours.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Les utilisateurs peuvent signaler des contenus ou des comptes qui enfreignent les règles de Medium ou envoyer un signalement décrivant l'infraction présumée. Les publications et les comptes signalés sont examinés par l'équipe de confiance et sécurité de Medium (Trust & Security), qui détermine si des règles ont été effectivement violées et, le cas échéant, prend les mesures nécessaires. Le coût économique marginal d'un recours à des modérateurs humains pour détecter les contenus répréhensibles est probablement relativement élevé. Medium n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	La violation des règles de Medium peut entraîner l'envoi d'avertissements, l'application de restrictions sur le compte, une limitation de la diffusion des publications et des contenus, la suspension des contenus et la suspension du compte en infraction. Les contenus controversés et extrémistes (sans qu'il soit fait référence spécifiquement à l'extrémisme violent) risquent particulièrement de faire l'objet d'une suspension ou d'une limitation de leur diffusion (Medium, n.d.).
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non. Medium a publié en 2015 (Medium, 2015) un rapport de transparence sur les demandes d'information ou de retrait de contenus émises par les autorités en 2014, mais celui-ci ne mentionne pas précisément les contenus terroristes et extrémistes violents.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et	Sans objet.

données figurant dans les rapports de transparence	
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Information inconnue.

39. Odnoklassniki

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition particulière des contenus terroristes et extrémistes violents. Les conditions d'utilisation d'Odnoklassniki interdisent toutefois la propagande ou l'apologie de la haine ou du suprémacisme sur la base de critères sociaux, raciaux, nationaux ou religieux, les contenus comprenant des menaces ou incitant à la violence ou à commettre des infractions pénales, et la publication d'informations à caractère extrémiste. Le terme « extrémiste » n'est pas défini.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://ok.ru/regulations.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Quelques-unes de ces règles sont données à l'adresse https://ok.ru/help/54/4532. Les utilisateurs peuvent utiliser OK Live de manière anonyme s'ils acceptent des restrictions de fonctionnalité. Pour pouvoir profiter de l'ensemble des fonctionnalités proposées, les utilisateurs doivent soit utiliser leur profil Odnoklassniki, soit créer un nouveau profil avec leur numéro de téléphone.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits et de procédures de recours contre ces décisions	Odnoklassniki indique de manière générale qu'il peut avertir, notifier ou informer les utilisateurs du non-respect de ses conditions d'utilisation. Ils doivent dans ce cas obligatoirement suivre les instructions qui leur sont données. Odnoklassniki explique aussi qu'il peut supprimer les contenus qui selon lui enfreignent ou peuvent enfreindre la législation en vigueur, ses conditions d'utilisation, portent préjudice ou sont susceptibles de la faire, ou menacent la sécurité de ses utilisateurs ou de tiers.

4.1 Notifications des suppressions	Odnoklassniki notifie les utilisateurs de leurs infractions aux conditions d'utilisation à sa seule discrétion.
4.2 Procédures de recours contre une décision de retrait	Aucune procédure de recours n'est indiquée.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Des utilisateurs peuvent devenir modérateurs des pages personnelles d'autres utilisateurs ou créer des groupes dont ils sont ensuite modérateurs. Ils sont dans ce cas tenus de modérer les contenus publiés sur ces pages et dans ces groupes. Ils peuvent aussi devenir modérateurs de vidéos et de photos en téléchargeant l'application de modération d'Odnoklassniki (Odnoklassniki, n.d.). Les utilisateurs peuvent signaler les contenus contraires aux conditions d'utilisation d'Odnoklassniki. L'équipe de la plateforme examine les signalements et décide des mesures à prendre. Le coût économique marginal d'un recours à des modérateurs humains employés pour détecter les contenus répréhensibles est probablement relativement élevé. Les utilisateurs modérateurs n'engendrent aucun coût pour Odnoklassniki. Odnoklassniki indique qu'il ne censure de manière automatisée ni les informations publiées dans les parties accessibles au public de sa plateforme ou dans les pages personnelles des utilisateurs ni les messages personnels de ceux-ci, et qu'il n'a pas la capacité technique de le faire. Il n'effectue pas non plus de modération en amont des informations et des contenus publiés par les utilisateurs. Odnoklassniki n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	En cas de violation de ses conditions d'utilisation, Odnoklassniki a le droit de suspendre, de restreindre ou de supprimer l'accès de l'utilisateur concerné à sa plateforme.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.

9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. Des contenus terroristes et extrémistes violents soutenant l'État islamiste ont été trouvés sur Odnoklassniki (Clifford & Powell, 2019)

40. Haokan Video

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. Les conditions d'utilisation de Haokan interdisent toutefois d'utiliser les services de la plateforme pour mener des activités illicites ou inappropriées, dont la diffusion d'actes de violence, d'homicide et de terrorisme. Le terme « terrorisme » n'est pas défini.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://haokan.baidu.com/videoui/page/about#agreement.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté	Haokan indique de manière générale qu'il se réserve le droit de bloquer ou de retirer des contenus à tout moment sans notification s'il juge que ceux-ci enfreignent ses règles.

(suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	
4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Aucune procédure de recours n'est indiquée.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Haokan ne communique pas d'informations à cet égard. Haokan n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	Haokan indique qu'en cas de violation de ses conditions d'utilisation, il est autorisé à résilier ou à restreindre l'accès de l'utilisateur concerné à son compte et à supprimer les contenus en infraction sans avertissement préalable.
7. Publication par le service de rapports de transparence sur les	Non.

contenus terroristes et extrémistes violents	
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Information inconnue.

41. Smule

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. Les règles de la communauté de Smule interdisent toutefois les contenus qui promeuvent le sectarisme, la discrimination, la haine, l'intolérance ou le racisme, qui sont haineux, insultants ou choquants, ou qui incitent à la violence.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.smule.com/en/s/communityguidelines et https://www.smule.com/en/termsofservice.
3. Présence de dispositions précises	Non.

applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Smule indique de manière générale qu'il n'effectue aucun filtrage de contenu en amont, mais se réserve le droit de retirer ou de supprimer tout contenu à sa seule discrétion, avec ou sans avertissement préalable, en particulier s'il ne respecte pas ses directives communautaires ou ses conditions d'utilisation. S'il trouve un « contenu répréhensible », il prend des mesures appropriées, telles que avertir l'utilisateur, suspendre ou résilier son compte, supprimer tous ses contenus ou signaler l'utilisateur aux autorités répressives, de manière directe ou indirecte.
4.1 Notifications des suppressions ou des autres décisions de sanction	Les notifications prennent la forme d'avertissements envoyés à la discrétion de Smule.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Aucune procédure de recours n'est indiquée.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Les utilisateurs peuvent signaler les contenus contraires à ses directives communautaires et conditions d'utilisation. Smule examine les contenus signalés par ses membres et peut les retirer s'ils sont jugés inappropriés ou dangereux pour la communauté ou s'ils ne respectent pas de quelque manière que ce soit ses directives communautaires ou conditions d'utilisation. Le coût économique marginal d'un recours à des modérateurs humains pour détecter les contenus répréhensibles est probablement relativement élevé. Smule n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.

6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	Si un utilisateur enfreint ses directives communautaires ou conditions d'utilisation, Smule peut lui envoyer un avertissement, supprimer le contenu inapproprié, résilier définitivement son compte, notifier les autorités répressives ou engager des poursuites à son encontre.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Information inconnue.

42. KaKaoTalk

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Suite à la mise à jour récente de ses règles de fonctionnement, Kakao interdit la « publication de contenus contraires à la dignité humaine, incitant à la violence et favorisant la discrimination et les préjugés sur la base de motifs fondés sur le lieu d'origine de la personne (pays et région), sa race, son apparence, son handicap ou sa maladie, son sexe, son identité de genre, son orientation sexuelle et d'autres critères liés à son identité. Dans son « engagement à mettre fin aux propos haineux en ligne », Kakao définit les propos haineux comme un « discours

	insultant ciblant une personne ou un groupe de personnes » et accompagné d'« actes de discrimination, d'une incitation à nourrir des préjugés, d'insultes et d'une exclusion sociale pour des motifs liés notamment au lieu d'origine de la personne (pays et région), à sa race, à son apparence, à son handicap ou sa maladie, à son sexe, à son identité de genre, à son orientation sexuelle et à d'autres critères liés à son identité ».
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.kakao.com/policy/oppolicy?lang=en (article 3, paragraphe 2, point 15).
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Kakao TV applique de larges et strictes restrictions sur la diffusion en direct de contenus illicites, violents ou haineux. Les administrateurs de Kakao TV vérifient en temps réel tous les contenus diffusés en direct et informent les utilisateurs du service que ses responsables peuvent interrompre sans délai les diffusions en ligne chaque fois que le contenu enfreint ses règles. En vertu des règles communautaires de Kakao, les contenus diffusés en direct sur Kakao TV sont également soumis à un principe de lutte contre la discrimination qui interdit toutes formes d'expression pouvant induire des discriminations ou de promotion d'opinions biaisées sur la base de stéréotypes. Kakao TV est le seul service de Kakao permettant à ses utilisateurs de diffuser des contenus en direct à destination du public.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	KakaoTalk indique de manière générale qu'en cas de violation de ses règles ou de la législation en vigueur, il peut mener des enquêtes sur les infractions, supprimer les publications concernées à titre provisoire ou définitif, ou restreindre tout ou partie de ses services provisoirement ou définitivement. Le caractère provisoire ou définitif de la restriction dépend du nombre cumulé d'infractions. Cependant, toute activité interdite en vertu de la législation ou des réglementations en vigueur entraîne une restriction immédiate et définitive des services, quel que soit le nombre cumulé d'infractions.
4.1 Notifications des suppressions ou des autres décisions de sanction	Les utilisateurs sont avertis dans les meilleurs délais des mesures de sanction mentionnées ci-dessus par courrier électronique ou par d'autres outils disponibles sur l'application, sauf lorsqu'il faut protéger de manière urgente d'autres utilisateurs.

4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Les utilisateurs peuvent faire appel des mesures prises. Ils sont informés de la décision définitive de KakaoTalk lorsque leur demande de recours a été examinée.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Les utilisateurs peuvent créer une « chaîne de story », en devenir « maîtres » et y inviter des « directeurs ». Les maîtres et les directeurs sont les administrateurs et les modérateurs des chaînes. Ils peuvent bloquer et signaler les utilisateurs et les contenus contraires aux règles de KaKaoTalk. Les utilisateurs peuvent également signaler les contenus contraires aux règles de l'application. L'équipe de KaKaoTalk examine les signalements et, le cas échéant, prend les mesures nécessaires. Les autorités de régulation sud-coréennes, telles que la National Policy Agency (NPA), les Communications Commissions et la Korean Communications Standards Commission (KCSC), peuvent demander la suppression des informations antisociales, violentes et illégales. KaKaoTalk peut par ailleurs appliquer des restrictions en cas d'activités interdites en vertu de ses propres règles ou qui enfreignent la législation ou les réglementations en vigueur, sans avoir reçu de signalements de la part des utilisateurs ou des autorités de régulation. Kakao surveille les contenus des chaînes de story ainsi que ceux des blogues et du réseau social à partir de mots clés en rapport avec les contenus terroristes et extrémistes violents et illégaux. KaKao TV, la plateforme de diffusion de vidéos en ligne, est également surveillée, y compris le contenu diffusé en direct. Si la surveillance détecte un contenu problématique, tel que du terrorisme ou de l'extrémisme violent, KaKao TV demande à l'utilisateur qui l'a mis en ligne de le modifier (c'est-à-dire de le retirer ou de le modifier). S'il n'est pas modifié dans un délai de trois jours, il est supprimé par les modérateurs, qui excluent en outre l'utilisateur à titre provisoire ou définitif en fonction du degré de violence du contenu et du nombre cumulé d'infractions commises par l'utilisateur. Cependant, s'il est jugé que le contenu nécessite une intervention urgente, les modérateurs sont autorisés à le supprimer immédiatement. Le coût économique marginal d'une utilisation des outils automatisés pour détecter les contenus discutables est probablement très faible (bien que les coûts fixes puissent être élevés), alors que celui d'un recours à des modérateurs humains pour la même fonction est probablement relativement élevé. Les utilisateurs modérateurs n'engendrent aucun coût pour KaKaoTalk. KaKaoTalk n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.

6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	En cas de violation de ses règles, KaKaoTalk peut envoyer un avertissement, supprimer les contenus en infraction et restreindre ses services à titre provisoire ou définitif selon le nombre cumulé d'infractions. Cependant, toute activité interdite en vertu de la législation ou des réglementations en vigueur entraîne une restriction des services immédiate et définitive, quel que soit le nombre cumulé d'infractions.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non. KaKaoTalk publie des rapports de transparence (Daum Kakao, n.d.) indiquant le nombre de demandes formulées par les autorités sud-coréennes pour accéder aux informations des utilisateurs et retirer des contenus en infraction à ses conditions d'utilisation ou d'autres règles, mais sans mentionner spécifiquement les contenus terroristes et extrémistes violents.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Les rapports de transparence sont publiés deux fois par an.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Information inconnue.

43. DeviantArt

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. Les conditions d'utilisation de DeviantArt prévoient toutefois que les commentaires exagérément agressifs ou insultants sont interdits. Les utilisateurs ne doivent pas utiliser DeviantArt à des fins illégales ou pour mettre en ligne, publier ou diffuser autrement des contenus illégaux, menaçants, nuisibles ou répréhensibles de quelque manière que ce soit.

2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://about.deviantart.com/policy/service/, https://about.deviantart.com/policy/etiquette/ et https://about.deviantart.com/policy/submission/.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Lorsqu'un contenu interdit (appelé « écart ») est signalé, son auteur peut recevoir une notification anonyme lui demandant si le contenu est, par exemple, un « contenu adulte », selon le motif du signalement. L'auteur a ainsi la possibilité d'intervenir et de remédier au problème. S'il n'intervient pas et que le contenu ne fait l'objet d'aucun nouveau signalement, l'équipe peut décider qu'il n'est pas nécessaire de le supprimer ou de lui apposer une mention et invalider le signalement. En revanche, si le nombre de signalements augmente, le personnel traite le contenu en priorité et prend les mesures nécessaires plus rapidement, en lui ajoutant une mention ou en le supprimant, ou en invalidant le signalement. Il convient de souligner que même si une notification est envoyée à l'utilisateur concerné, les signalements sont systématiquement transmis au personnel de DeviantArt pour approbation. L'envoi de la notification permet simplement de laisser à l'utilisateur la possibilité de corriger une éventuelle erreur commise de bonne foi (Kitsune, 2017). L'utilisation des outils de communication fournis par DeviantArt à des fins délibérément agressives ou inappropriées peut entraîner l'application de mesures disciplinaires (DeviantArt, n.d.). Les fils de discussion des forums qui contiennent des propos déplacés, portent sur des sujets inappropriés ou comprennent un nombre indésirable d'infractions aux règles de DeviantArt sont bloqués et fermés aux commentaires. Un utilisateur inscrit sur DeviantArt peut être administrateur ou membre d'un « groupe », c'est-à-dire un ensemble d'applications et de pages d'utilisateurs créé pour rassembler des contenus, des discussions et des membres autour d'intérêts communs. Les administrateurs peuvent définir leurs propres règles et les droits accordés aux utilisateurs ou membres de leur groupe. D'une manière générale, DeviantArt n'intervient pas dans les groupes, sauf en cas de violation manifeste de ses règles. Il peut alors retirer le groupe concerné et les droits qui lui sont associés.

	Les utilisateurs qui ont fait preuve d'un comportement inacceptable en ne respectant pas les règles de DeviantArt ou en se livrant à des activités inappropriées ou préjudiciables à la communauté peuvent faire l'objet d'une suspension provisoire de leur compte (DeviantArt, n.d.). Si un compte est suspendu, le message « compte suspendu » s'affiche pendant toute la durée de la suspension à la place de la page de profil habituelle de l'utilisateur. La durée des suspensions administratives peut varier. Elles sont le plus souvent imposées pour 24 heures, une (1) semaine, deux (2) semaines ou trente (30) jours (un mois). Pendant ce temps, l'utilisateur ne peut plus effectuer de publications, utiliser la plupart des fonctionnalités de la plateforme ou interagir avec la communauté en général. L'utilisateur reçoit une notification de la mesure prise, qui peut être assortie d'un message privé ou du motif de la mesure, et un minuteur est ajouté sur la page de son profil. Si une mesure disciplinaire doit être prise, celle-ci tient compte des suspensions précédemment enregistrées pour la personne concernée et peut entraîner, le cas échéant, une suspension plus longue ou, si la personne s'est déjà signalée plusieurs fois, la résiliation de son compte (DeviantArt, n.d.).
4.1 Notifications des suppressions ou des autres décisions de sanction	Si un contenu est supprimé par le personnel de DeviantArt, son auteur reçoit une notification. Les suspensions de compte sont également notifiées.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Si l'auteur pense que son contenu est autorisé et que le personnel a commis une erreur, il peut contester la décision en expliquant ses raisons. Le personnel examinera alors une nouvelle fois le contenu. DeviantArt autorise généralement ses utilisateurs à déposer un recours et à demander des précisions sur des suppressions de contenu, des avertissements pour infractions, des suspensions et résiliations de compte ou toute autre mesure administrative. Les recours, demandes de renseignements et questions sont examinés et traités par le personnel de DeviantArt.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Les administrateurs de groupe sont modérateurs au sein de leur groupe. Les utilisateurs peuvent également signaler les contenus contraires aux règles de l'application. Lorsque le personnel de DeviantArt est informé d'une infraction, il examine le signalement et, le cas échéant, prend les mesures nécessaires. DeviantArt indique ne pas avoir la possibilité de contrôler les contenus que les utilisateurs peuvent mettre en ligne, publier ou diffuser sur sa plateforme et ne pas être soumis à l'obligation de surveiller ces contenus pour quelque raison que ce soit. Le coût économique marginal d'un recours à des modérateurs humains employés pour détecter les contenus répréhensibles

	est probablement relativement élevé. Les utilisateurs modérateurs n'engendrent aucun coût pour DeviantArt. DeviantArt n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	La violation des règles de DeviantArt peut entraîner l'envoi d'un avertissement, la suppression du contenu, la suspension du compte ou la résiliation de l'adhésion de l'utilisateur, à la discrétion exclusive de DeviantArt.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. Des groupes néonazis ont utilisé DeviantArt pour mettre en ligne de la propagande et recruter des membres (Hayden, Mysterious Neo-Nazi Advocated Terrorism for Six Years Before Disappearance, 2019).

44. Meetup

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. Cependant, au titre des conditions d'utilisation de Meetup, il est interdit de publier des contenus violents explicites. Les comportements incitant à la violence envers des individus ou des groupes d'individus, sur la simple base de leurs origines ou de leurs croyances sont interdits. De plus, il est interdit d'utiliser Meetup pour promouvoir, faciliter ou organiser des activités à caractère violent, criminel ou non consenties qui mettraient en péril autrui, que ce soit physiquement, psychologiquement ou émotionnellement. Par ailleurs les « groupes » (espaces créés autour d'intérêts ou d'activités précis) ne doivent pas contenir ou promouvoir des événements qui organisent, encouragent, proposent ou diffusent des services ou recrutent pour des organisations terroristes, comprendre des contenus ou promouvoir des événements susceptibles de menacer la sécurité du public ou des personnes, tels que des encouragements, des incitations ou des déclarations intentionnelles ou des menaces de commettre un acte de violence à l'encontre d'un groupe, d'une personne ou d'un lieu, faire l'apologie des armes et de la fabrication d'explosifs, et inciter à la violence en réaction à des événements privés ou publics.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://help.meetup.com/hc/en-us/articles/360002897532-Usage-and-content-policies-Rules-for-using-Meetup, https://help.meetup.com/hc/fr-fr/articles/360002897532-Politique-d-utilisation-et-de-contenu, https://help.meetup.com/hc/fr-fr/articles/360002897712-Politique-concernant-les-groupes-Meetup-normes-et-standards- et https://help.meetup.com/hc/fr-fr/articles/360027447252-Conditions-g%C3%A9n%C3%A9rales-de-service.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de	Meetup prévoit qu'une violation de ses règles et conditions d'utilisation peut entraîner la modification, la suspension ou la résiliation du compte ou de l'accès à la plateforme. Le cas échéant, il informe l'utilisateur concerné des motifs de la modification, de la suspension ou de la résiliation.

notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	
4.1 Notifications des suppressions ou des autres décisions de sanction	Les utilisateurs sont informés des décisions de sanction.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Si un utilisateur pense que son compte a été modifié, suspendu ou résilié par erreur, il peut faire appel de la décision.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Les administrateurs des groupes modèrent les contenus de leurs groupes et peuvent modifier, suspendre ou résilier l'accès des utilisateurs à ces groupes. Les utilisateurs peuvent également signaler les contenus contraires aux règles de la plateforme. L'équipe de confiance et de sécurité de Meetup examine les signalements reçus et, le cas échéant, prend les mesures nécessaires. Le coût économique marginal d'un recours à des modérateurs humains employés pour détecter les contenus répréhensibles est probablement relativement élevé. Les utilisateurs modérateurs n'engendrent aucun coût pour Meetup. Meetup indique qu'il n'examine généralement pas les contenus avant qu'ils soient publiés (Meetup, 2019). Meetup n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	La violation des règles de Meetup peut entraîner la suppression du contenu et la modification, la suspension ou la résiliation du compte de l'utilisateur concerné.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non. Meetup publie des rapports de transparence (Meetup, 2017) qui révèlent les demandes d'accès aux informations des utilisateurs émanant des autorités et les demandes de retrait de contenu en raison d'atteinte aux droits de la propriété intellectuelle, mais qui ne mentionnent pas les contenus terroristes et extrémistes violents.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.

9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Information inconnue.

45. 4chan

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. Les conditions d'utilisation de 4chan interdisent toutefois les contenus qui enfreignent la législation locale ou celle des États-Unis.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur http://www.4chan.org/rules#global4.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de	Selon 4chan, les fils de discussion arrivent à expiration ou sont supprimés par le logiciel de la plateforme assez rapidement. La plupart des tableaux (thématiques de fils de discussion) étant limités à dix pages, les contenus ne peuvent généralement être vus que pendant quelques heures ou quelques jours avant d'être retirés. Les publications sont la plupart du temps retirées automatiquement, mais ils peuvent parfois l'être par un modérateur ou « gardien ». Les modérateurs sont des personnes choisies pour assurer la maintenance générale de la plateforme. Ils peuvent supprimer des publications publiées sur toute la plateforme, exclure des

procédures de recours contre ces décisions	utilisateurs, fermer des fils de discussion et prendre des mesures connexes. Les gardiens constituent une catégorie située entre l'utilisateur et le modérateur. Ils ont accès au système de signalement de la plateforme et peuvent à la fois supprimer des publications sur les tableaux qui leur sont confiés et introduire des demandes d'exclusion. Ils sont sélectionnés sur candidature à l'issue d'un processus de tests et d'intégration. L'entrée dans l'équipe de modération se fait uniquement sur invitation. Le programme des gardiens accueille de temps à autre de nouveaux participants. Il n'est conservé aucun enregistrement public des contenus supprimés et, les fils de discussion étant retirés régulièrement, il n'existe aucun moyen de savoir quels contenus ont été retirés par l'équipe de modération. Autrement dit, il est impossible pour l'utilisateur de savoir précisément sur quels contenus porte la modération ni à quel moment elle est effectuée. Les modérateurs de 4chan se réservent le droit de bloquer ou d'interdire l'accès d'un utilisateur, ou de supprimer des contenus pour quelque raison que ce soit sans avertissement préalable. Les utilisateurs ne peuvent plus temporairement publier de publications lorsqu'une demande d'exclusion de leur adresse IP est en attente de traitement. Le blocage dure 15 minutes à partir du moment où un gardien introduit la demande d'exclusion et est levé immédiatement si un modérateur rejette la demande. Si en revanche celle-ci est approuvée, l'exclusion est appliquée.
4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	S'ils pensent qu'une erreur a été commise, les utilisateurs peuvent déposer un recours en contactant les modérateurs.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage	4chan indique qu'il encourage le signalement des publications (4chan, n.d.). Celles-ci sont ensuite examinées par les modérateurs qui, le cas échéant, prennent les mesures nécessaires. Le coût économique marginal d'un recours à des modérateurs humains employés pour détecter les contenus répréhensibles est probablement relativement élevé. Les utilisateurs modérateurs n'engendrent aucun coût pour 4chan.

d'empreintes numériques ou d'adresses URL)	4chan n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	La violation des règles de 4chan peut entraîner la suppression de la publication ainsi qu'une exclusion temporaire voire, dans certains cas, définitive.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. La propagande néonazie, par exemple, est fréquente sur 4chan (Arthur, 2019).

46. Google Drive

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition particulière des contenus terroristes et extrémistes violents. Cependant, le règlement du programme concernant l'utilisation abusive de Google (Google, n.d.), qui s'applique aussi à Google Drive, comprend des dispositions particulières sur la violence, les incitations à la haine et les activités à caractère terroriste. *Violence* : il est interdit de menacer une personne de violences graves ou de mort, ou d'obtenir l'appui d'autres personnes dans

	le but de lui porter physiquement atteinte. En cas de menace grave et imminente de blessure physique ou de mort, Google est susceptible de prendre les mesures nécessaires concernant le contenu. La violence gratuite ou les contenus à caractère sanglant visant à choquer ou à susciter un intérêt malsain ne sont pas autorisés. S'ils publient des contenus explicites dans un contexte documentaire, scientifique, artistique ou en lien avec l'actualité, les utilisateurs doivent donner suffisamment d'informations pour que les autres personnes comprennent ce dont il s'agit. Dans certains cas, les contenus publiés peuvent être si violents ou choquants qu'aucun contexte ne pourra justifier leur présence sur les plateformes de Google. Les utilisateurs ne doivent pas encourager les autres personnes à commettre des actes de violence spécifiques. *Incitation à la haine* : l'incitation à la haine est interdite. L'incitation à la haine désigne tout contenu qui incite à la violence ou la justifie, ou dont l'objectif principal est d'inciter à la haine envers une personne ou un groupe en raison de son origine ethnique, de sa religion, de son handicap, de son âge, de sa nationalité, de son statut d'ancien combattant, de son sexe, de son orientation ou identité sexuelle ou de toute autre caractéristique associée à une discrimination ou une à marginalisation systématiques. *Activités à caractère terroriste* : les organisations terroristes ne sont pas autorisées à utiliser Google Drive à quelque fin que ce soit, y compris pour le recrutement. Les contenus à caractère terroriste, tels que la promotion d'actes terroristes, l'incitation à la violence ou l'apologie d'attentats terroristes, sont aussi formellement interdits. Le terme « organisation terroriste » n'est pas défini. Si les utilisateurs publient des contenus liés au terrorisme dans un objectif pédagogique, documentaire, scientifique ou artistique, ils doivent donner suffisamment d'informations pour que les autres personnes comprennent le contexte.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.google.com/drive/terms-of-service/ et https://support.google.com/docs/answer/148505?visit_id=637064013896463652-1393240150&rd=1.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Non.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des	Si des fichiers font l'objet d'un signalement pour infraction, un drapeau peut s'afficher à côté de leur nom et leur propriétaire ne peut plus les partager. Les fichiers ne sont plus accessibles publiquement, même pour les utilisateurs qui disposent du lien.

règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	S'ils pensent que leurs fichiers n'enfreignent pas les conditions d'utilisation ou le règlement du programme Google Drive, ils peuvent demander qu'ils soient examinés (Google, n.d.). Si un utilisateur enfreint de façon significative ou répétée les conditions d'utilisation ou le règlement du programme, Google est susceptible d'interrompre ou de désactiver définitivement son accès à Google Drive. Dans ce cas, l'utilisateur en est informé à l'avance. Google peut toutefois être amené à suspendre ou à désactiver l'accès d'un utilisateur à Google Drive sans préavis s'il utilise ce service d'une façon susceptible de mettre en jeu la responsabilité légale de Google ou d'empêcher d'autres personnes d'utiliser Google Drive.
4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Aucune procédure de recours n'est indiquée.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Les utilisateurs peuvent signaler les contenus contraires aux conditions d'utilisation et au règlement de Google Drive. Les signalements sont examinés par le personnel de Google. Google indique que l'envoi d'un signalement ne garantit pas que le fichier concerné sera supprimé ni qu'il prendra des mesures en conséquence. Les contenus qui déplaisent à un utilisateur ou qui lui semblent inappropriés n'enfreignent en effet pas forcément les conditions d'utilisation ou le règlement du programme. Google précise également qu'il est susceptible d'examiner les contenus et le comportement des utilisateurs sur Google Drive pour vérifier qu'ils respectent les conditions d'utilisation et le règlement du programme (Google, 2019). La société a indiqué que les fichiers stockés sur Google Drive sont analysés par un algorithme qui détecte les infractions à son règlement et bloque automatiquement les fichiers présumés ne pas le respecter. Ce système n'implique aucune évaluation humaine (Titcomb, 2017). Le coût économique marginal d'une utilisation des outils automatisés pour détecter les contenus discutables est probablement très faible (bien que les coûts fixes puissent être élevés), alors que celui d'un recours à des modérateurs humains pour la même fonction est probablement relativement élevé. Google Drive n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions	En cas de mise en ligne d'un contenu inapproprié contraire aux conditions d'utilisation ou aux autres règlements, Google peut prendre les mesures suivantes :

d'utilisation ou des règles de la communauté	- retirer le fichier du compte ; - appliquer des restrictions au partage du fichier ; - limiter l'accès au fichier ; - désactiver l'accès à un ou plusieurs produits Google ; - supprimer le compte Google. (Google, n.d.)
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non. Google publie des rapports de transparence (Google, n.d.) couvrant l'ensemble des produits et services de Google, y compris Google Drive. Les rapports comportent une section sur les demandes de retrait de contenu émanant des autorités en raison d'infractions avec les législations locales ou les conditions d'utilisation ou règlements de Google, mais ne mentionnent pas spécifiquement les contenus terroristes et extrémistes violents.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. Des comptes de l'EILL ont été découverts sur Google Drive (Katz, To Curb Terrorist Propaganda Online, Look to YouTube. No, Really., 2018).

47. Dropbox

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. La politique d'utilisation acceptable de Dropbox stipule toutefois que les utilisateurs ne doivent pas publier ou partager du contenu contenant des actes de violence extrême ou des actes terroristes, notamment de propagande terroriste. L'utilisation du service pour inciter au sectarisme ou à la haine envers une personne ou un groupe de personnes en raison de sa religion, de son origine ethnique, de son sexe, de son identité de genre, de son orientation sexuelle, d'un handicap ou d'une déficience est également interdite.

2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.dropbox.com/terms et https://www.dropbox.com/terms#acceptable_use.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Sans objet.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Dropbox indique qu'il est en droit de suspendre ou de résilier l'accès d'un utilisateur à ses services s'il ne respecte pas ses conditions d'utilisation ou s'il utilise les services d'une manière susceptible de causer un risque réel de dommages ou de pertes pour Dropbox ou les autres utilisateurs. Pour le cas où il enverrait un avertissement préalable à l'utilisateur, Dropbox donne à celui-ci l'occasion d'exporter ses contenus. Si, après réception de cette notification, l'utilisateur ne prend pas les mesures demandées, Dropbox suspendra ou résiliera son accès aux services. Dropbox n'envoie pas de notification à l'avance si l'utilisateur commet une violation substantielle des conditions d'utilisation, à savoir une violation qui risquerait d'engager la responsabilité de Dropbox ou de compromettre sa capacité à fournir ses services aux autres utilisateurs, ou si la loi le lui interdit.
4.1 Notifications des suppressions ou des autres décisions de sanction	Aucune notification n'est indiquée.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Les utilisateurs peuvent demander à Dropbox de revenir sur la décision de retrait s'ils pensent que les contenus concernés n'enfreignent pas les conditions d'utilisation de la plateforme.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Les utilisateurs, de même que des tiers, y compris des signaleurs de confiance et des organisations non gouvernementales, peuvent signaler les contenus contraires aux conditions d'utilisation et aux règles de Dropbox. L'équipe de Dropbox examine les signalements, mène une enquête sur l'infraction présumée et, le cas échéant, prend les mesures nécessaires. Dropbox recourt également à des outils de détection automatique et emploie une équipe d'évaluateurs humains. La plateforme a indiqué que son personnel avait besoin en de rares occasions d'accéder au contenu des fichiers des utilisateurs, en particulier pour faire appliquer ses conditions d'utilisation et ses règles (Dropbox, n.d.). Le coût économique marginal d'un recours à des modérateurs humains pour détecter les contenus répréhensibles est probablement relativement élevé. Dropbox est membre du GIFCT.

6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	La violation des conditions d'utilisation ou des règles de Dropbox peut entraîner la perte du droit d'accéder aux services de la plateforme pour l'utilisateur, ou la suspension ou la résiliation de son compte.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non. Dropbox publie des rapports de transparence (Dropbox, n.d.) dont une section porte sur les demandes de retrait de contenu émanant des autorités en raison d'infractions avec les législations locales ou les conditions d'utilisation ou règlements de Dropbox, mais ceux-ci ne mentionnent pas spécifiquement les contenus terroristes et extrémistes violents.
8. Informations ou types de données figurant dans les rapports de transparence	Sans objet.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Sans objet.
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. Du contenu de l'EILL a été découvert sur Dropbox (Bennett, 2019).

48. Microsoft OneDrive

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. Le contrat de services de Microsoft, qui régit OneDrive, interdit toute activité nuisible à d'autres personnes, telle que la publication de contenu terroriste ou extrémiste violent, des propos haineux ou des appels à la violence contre des tiers. Microsoft indique que, dans le cadre de ses services, un contenu terroriste désigne un contenu publié par une organisation figurant sur la liste récapitulative du Conseil de sécurité des Nations unies (Conseil de Sécurité des Nations Unies), ou visant à soutenir une telle organisation, qui représente explicitement la violence, encourage les actes violents, cautionne une organisation terroriste ou ses actes, et incite à rejoindre ces groupes. La liste récapitulative du Conseil de sécurité des Nations Unies répertorie les groupes considérés par le Conseil de sécurité des Nations unies comme des organisations terroristes (Microsoft, 2016).

	Dans son Digital Safety Content Report (Microsoft, 2021), Microsoft explique clairement que les « contenus tant terroristes qu'extrémistes violents sont interdits sur les plateformes et services Microsoft » et que le code de conduite contractuel des services Microsoft (Microsoft Services Agreement Code of Conduct) interdit la « publication de contenu terroriste ou extrémiste violent ».
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://www.microsoft.com/en-us/servicesagreement/.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Sans objet.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	Microsoft indique qu'il se réserve le droit de retirer ou de bloquer le contenu d'un utilisateur sur OneDrive à quelque moment que ce soit s'il est porté à sa connaissance que le contenu est susceptible d'enfreindre la législation applicable ou son contrat de services. Lors des enquêtes relatives aux infractions présumées au contrat de services, Microsoft se réserve le droit de consulter le contenu afin de résoudre le problème. La société précise toutefois qu'elle ne surveille pas OneDrive. Elle suit une procédure « de notification et de retrait » pour le retrait des contenus interdits, dont les contenus terroristes. Selon cette procédure, Microsoft reçoit une « notification » (de la part des autorités ou d'un utilisateur, par exemple), puis elle retire le contenu concerné. Si la présence de contenus terroristes sur les services aux consommateurs qu'elle héberge, dont OneDrive, est portée à sa connaissance par ses outils de signalement en ligne, elle les supprime (Microsoft, 2016). Ainsi que le prévoit le contrat de services de Microsoft, « si vous enfreignez (…) les présentes conditions, nous pouvons (...) cesser de vous fournir les services ou fermer votre compte Microsoft. Nous nous réservons également le droit de supprimer ou de bloquer votre contenu des services à tout moment si nous pensons qu'il pourrait enfreindre la réglementation applicable ou les présentes conditions. Lors des enquêtes relatives aux infractions suspectées des présentes conditions, Microsoft se réserve le droit de consulter votre contenu afin de résoudre le problème. »

4.1 Notifications des suppressions ou des autres décisions de sanction	Les notifications sont envoyées à la discrétion de Microsoft. Selon le contrat de services Microsoft, « si une information doit vous être communiquée concernant un service que vous utilisez, nous vous enverrons les notifications de service (...). Si vous nous avez donné votre adresse e-mail ou votre numéro de téléphone dans le cadre de votre compte Microsoft, vous êtes susceptible de recevoir des notifications de service par e-mail ou SMS, y compris pour vérifier votre identité avant d'enregistrer votre numéro de téléphone mobile et de vérifier vos achats. Vous êtes susceptible de recevoir des notifications de service par d'autres moyens (par exemple, par des messages intégrés au produit). »
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Le formulaire pour faire appel de la suspension d'un compte Microsoft est disponible à l'adresse https://www.microsoft.com/en-us/concern/AccountReinstatement.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	Microsoft indique que le code de conduite du contrat de services Microsoft interdit la « publication de contenu terroriste ou extrémiste violent ». Microsoft encourage le signalement de contenu publié par une organisation terroriste, ou visant à soutenir une telle organisation, qui représente explicitement la violence, encourage les actes violents, cautionne une organisation terroriste ou ses actes, et incite à rejoindre ces groupes. Microsoft évalue ces rapports, prend les mesures nécessaires concernant les contenus et, le cas échéant, suspend les comptes associés à des infractions à son code de conduite. En outre, Microsoft met en œuvre différents outils pour détecter les contenus terroristes et extrémistes violents, dont une technologie de comparaison d'empreintes numériques et d'autres formes de détection proactive. Microsoft recourt à des outils d'analyse et de reconnaissance (telles que PhotoDNA ou MD5) ainsi qu'à d'autres solutions faisant appel à l'intelligence artificielle, notamment des programmes de catégorisation de textes et d'images ainsi que des techniques de détection de toilettage pour détecter les contenus terroristes et extrémistes violents. (Microsoft, 2021) Le coût économique marginal d'un recours à des modérateurs humains pour détecter les contenus répréhensibles est probablement relativement élevé. Microsoft est un membre fondateur du GIFCT et participe au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions	Si un utilisateur publie un contenu interdit ou qui enfreint de manière substantielle le contrat de services, Microsoft est susceptible de prendre des mesures à son encontre, telles qu'interrompre l'accès à OneDrive, fermer immédiatement le

d'utilisation ou des règles de la communauté	compte Microsoft de l'utilisateur ou bloquer l'envoi des communications (comme le courrier électronique, le partage de fichiers ou la messagerie instantanée) vers ou depuis OneDrive. Microsoft peut également bloquer ou retirer le contenu concerné. Voir aussi la section 4 ci-dessus ainsi que l'article de blogue publié en 2016, ci-dessous : « **Application de la procédure de notification et retrait** : nous continuons à appliquer la procédure de notification et de retrait pour supprimer les contenus interdits, dont les contenus terroristes. Si la présence de contenus terroristes sur les services aux consommateurs que nous hébergeons est portée à notre connaissance par nos outils de signalement en ligne, nous les supprimons. Tous les signalements de contenu terroriste - effectués par les autorités, des citoyens ou d'autres groupes - publié sur un service Microsoft doivent nous être envoyés par l'intermédiaire de ce formulaire. » (https://blogs.microsoft.com/on-the-issues/2016/05/20/microsofts-approach-terrorist-content-online/)
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Oui. Les chiffres concernant le nombre de contenus terroristes et extrémistes violents pour Skype figurent dans le Digital Safety Content Report de Microsoft (Microsoft, 2021). Ce rapport couvre les produits et services de Microsoft destinés au public, notamment OneDrive, Outlook, Skype, Bing et Xbox. Il convient de noter que les chiffres relatifs aux contenus terroristes et extrémistes violents sont rapportés collectivement pour tous les produits et services de Microsoft destinés au public et non par produit.
8. Informations ou types de données figurant dans les rapports de transparence	• Nombre de contenus terroristes et extrémistes violents ayant fait l'objet de mesures • Nombre de comptes suspendus en raison de contenus terroristes et extrémistes violents • Pourcentage de contenus terroristes et extrémistes violents détectés par Microsoft • Pourcentage de comptes suspendus en raison de contenus terroristes et extrémistes violents qui ont été rétablis après recours
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Le terme « contenu ayant fait l'objet de mesures » (*content actioned*) désigne un élément de contenu publié par un utilisateur que Microsoft a retiré de ses produits et services ou que Microsoft a bloqué pour empêcher les utilisateurs d'y accéder. Le terme « suspension de compte » (*account suspension*) signifie retirer à l'utilisateur la possibilité d'accéder au compte du service de manière soit permanente, soit temporaire.

	Le terme « détection proactive » (*proactive detection*) indique que le signalement d'un contenu sur les produits ou services est le fait de Microsoft, que ce soit de manière automatisée ou par évaluation manuelle.
10. Fréquence de publication des rapports de transparence	Cette information n'est pas mentionnée.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. Des vidéos de l'EILL ont été hébergées sur OneDrive (Counter Extremism Project, 2018).

49. WordPress.com

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et violents, mais les conditions d'utilisation de WordPress.com interdisent les sites internet des groupes terroristes reconnus comme tels par les autorités américaines. L'Office of Foreign Assets Control (OFAC) du Trésor américain tient à jour la liste des « ressortissants spécialement désignés » (US Treasury, 2020), avec lesquels WordPress.com n'a pas le droit de faire des affaires. WordPress.com interdit aux personnes, groupes ou entités figurant sur cette liste d'utiliser ses services (WordPress, n.d.). Les incitations explicites à la violence sont également interdites. Cela comprend la publication de contenus qui profèrent des menaces, encouragent ou incitent à la violence ou à causer des préjudices physiques ou la mort, menacent des personnes ou des groupes ciblés ou commettent d'autres actes de violence aveugles.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://en-gb.wordpress.com/tos/ et https://en.support.wordpress.com/user-guidelines/.
3. Présence de dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	Sans objet.
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de	WordPress.com a travaillé en collaboration avec des spécialistes de l'extrémisme en ligne et les autorités répressives pour élaborer des mesures destinées à lutter contre la propagande terroriste et extrémiste (pas spécifiquement extrémiste violente). La plateforme peut

contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	suspendre les sites internet qui incitent à la violence ou qui sont liés à des groupes terroristes officiellement interdits (au titre de la liste de l'OFAC du Trésor américain), quel que soit leur contenu. Elle applique également d'autres mesures. Elle peut par exemple signaler un contenu ou retirer un site du Lecteur WordPress.com pour que son contenu soit plus difficile à trouver. Un site qui fait l'objet d'un signalement est automatiquement retiré de tous les programmes publicitaires gérés par WordPress.com. D'après la plateforme, les signalements effectués par les unités de référence internet (IRU) publiques représentent un moyen important d'attirer son attention sur les sites extrémistes (mais non spécifiquement extrémistes violents, rappelons-le). Ces organismes possèdent en effet une expertise de la propagande en ligne dont ne peuvent pas se doter les sociétés de technologie privées. Ils s'emploient à détecter les sites utilisés par des terroristes connus pour diffuser de la propagande ou organiser des actes de violence. Ils signalent les sites terroristes à WordPress.com par le biais d'une adresse électronique spéciale qui permet à la société de repérer plus facilement les signalements provenant d'une source fiable. WordPress.com ne retire pas automatiquement les sites internet de sa plateforme. Un membre de son équipe Risque et sécurité examine tous les signalements reçus pour déterminer si les contenus concernés enfreignent les règles de la plateforme. L'une des raisons d'examiner tous les signalements est d'éviter de retirer des contenus postés sur des sites légitimes (organes d'informations, sites universitaires) qui parlent du terrorisme ou de groupes terroristes. WordPress.com héberge les sites d'un certain nombre de très grands organes d'informations, de blogueurs d'actualité, d'universitaires et de chercheurs, qui publient tous des contenus légitimes sur le terrorisme. Dans un autre contexte, certains de ces contenus pourraient toutefois être considérés comme de la propagande terroriste, auquel cas ils seraient retirés de la plateforme au titre des règles qu'elle applique. WordPress.com indique que le contexte est très important et qu'elle ne peut confier à un robot les décisions susceptibles d'affecter un contenu légitime. Dans la mesure où le volume des signalements qu'elle reçoit n'est pas très important comparé à d'autres plateformes, elle peut effectuer plus d'examens humains que d'examens automatisés (Clicky, 2017).
4.1 Notifications des suppressions ou des autres décisions de sanction	WordPress.com explique que, en fonction du scénario, elle enverra un courrier électronique ou ajoutera une notification d'avertissement au tableau de bord de l'utilisateur qui a enfreint ses règles. La notification contiendra un lien que l'utilisateur pourra utiliser pour contacter la plateforme au sujet du problème. Wordpress.com ne précise toutefois pas ce que sont ces « scénarios » (WordPress.com, n.d.).

4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Les utilisateurs peuvent faire appel des mesures de sanction appliquées par WordPress.com s'ils estiment qu'elles ont été prises par erreur. La demande sera étudiée par une personne réelle et l'utilisateur sera informé de sa décision dès que possible.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel], bases de données de partage d'empreintes numériques ou d'adresses URL)	WordPress.com n'effectue pas de filtrage en amont des contenus publiés par les utilisateurs. Les utilisateurs peuvent signaler les contenus ou les sites qu'ils estiment contraires aux règles de WordPress.com. En outre, comme cela a été indiqué plus haut, les unités de référence internet (IRU) signalent les sites terroristes et extrémistes à la plateforme. Celle-ci les examine et, le cas échéant, prend les mesures nécessaires. Le coût économique marginal d'un recours à des modérateurs humains pour détecter les contenus répréhensibles est probablement relativement élevé. WordPress.com n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	Si WordPress.com détecte un site ou un contenu contraire à ses règles, elle supprime le contenu, désactive certaines fonctionnalités du compte ou suspend la totalité du site.
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Oui. Automattic, la société mère de WordPress.com, publie des rapports de transparence dont une section porte sur les signalements effectués par les unités de référence internet (IRU) pour des contenus extrémistes (mais pas spécifiquement extrémistes violents) (Automattic, n.d.). Le dernier rapport comprend les données du 1er janvier au 30 juin 2020.
8. Informations ou types de données figurant dans les rapports de transparence	- Nombre de notifications de contenu extrémiste (pas spécifiquement extrémiste violent) des unités de référence internet (IRU) - Nombre de notifications qui ont entraîné un retrait du contenu ou du site - Pourcentage des notifications qui ont entraîné un retrait du contenu ou du site Les chiffres sont répartis par mois (de janvier à juin et de juillet à décembre) et entité ou pays de signalement. Dans le résumé de son rapport de transparence, Automattic rend également compte du nombre de sites ou de contenus spécifiés dans les avis des unités de référence internet pour la période du 1er janvier 2018 au 30 juin 2020.

9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans les rapports de transparence	Aucune information n'est communiquée à cet égard.
10. Fréquence de publication des rapports de transparence	Les rapports font l'objet d'une édition semestrielle. Automattic a publié des rapports de transparence pour les périodes suivantes : - 2017 : 1er juillet - 31 décembre - 2018 : 1er janvier - 30 juin - 2018 : 1er juillet - 31 décembre - 2019 : 1er janvier - 30 juin - 2019 : 1er juillet - 31 décembre - 2020 : 1er janvier - 30 juin
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Oui. Voir la section 7 ci-dessus.

50. Wikipédia

1. Définition des contenus terroristes et extrémistes violents dans les conditions d'utilisation, les lignes de conduite ou les règles de la communauté	Il n'existe pas de définition des contenus terroristes et extrémistes violents. Les conditions d'utilisation de la Fondation Wikimedia, qui régissent Wikipédia, interdisent toutefois notamment le harcèlement, les menaces, les propos outrageants et le vandalisme. Il est également interdit d'utiliser les services de Wikimedia d'une façon incompatible avec la loi applicable.
2. Manière dont les conditions d'utilisation ou les règles de la communauté sont communiquées	Les textes sont disponibles sur https://foundation.wikimedia.org/wiki/Terms_of_Use/en et https://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines#Enforcement.
3. Présence de	Sans objet.

dispositions précises applicables aux contenus diffusés en direct dans les conditions d'utilisation ou les règles de la communauté	
4. Politiques et procédures de mise en œuvre et d'application des conditions d'utilisation ou des règles de la communauté (suppression de contenus) : en particulier, existence ou non de notifications des retraits ou d'autres décisions de sanction et de procédures de recours contre ces décisions	La communauté Wikipédia occupe un rôle de premier plan dans la définition et l'application des règles. Elle se compose des fonctions suivantes : - *Contributeurs* : bénévoles qui écrivent et modifient les pages de Wikipédia. - *Stewards* : contributeurs bénévoles chargés de la mise en œuvre technique des droits des utilisateurs. Ils sont habilités à utiliser la fonction Checkuser (Wikipédia, 2019) et à masquer les contenus (Wikipédia, 2020). - *Bureaucrates* : contributeurs volontaires habilités à accorder ou retirer à d'autres utilisateurs le statut d'administrateur ou de bureaucrate, et à attribuer ou supprimer le statut de bot à un compte. - *Administrateurs* : contributeurs auxquels est confié l'accès à des fonctionnalités techniques particulières (outils). Ils peuvent par exemple protéger et supprimer des pages ou bloquer des contributeurs (Wikipédia, 2020). Les principales règles de Wikipédia relatives aux contenus sont les suivantes : 1. Neutralité de point de vue : tous les articles de Wikipédia et les contenus encyclopédiques doivent être rédigés avec neutralité, en présentant les principaux points de vue de manière impartiale et proportionnelle, sans parti pris. 2. Vérifiabilité : les personnes qui lisent et modifient l'encyclopédie peuvent vérifier que les informations proviennent d'une source fiable. 3. Pas de publication de travaux inédits : Wikipédia ne publie pas de recherches qui n'ont encore jamais été publiées. Tous les contenus publiés sur Wikipédia doivent être associés à des sources fiables et publiées (Wikipédia, 2019). Les contenus sont supprimés par les administrateurs s'ils estiment qu'ils enfreignent les règles de contenu ou les autres règles de Wikipédia ou la législation des États-Unis (Wikipédia, 2020). La suppression repose sur les processus appliqués lors de la mise en œuvre et de l'enregistrement des décisions de la communauté concernant la suppression de pages et de médias (Wikipedia, 2020). Une discussion doit en principe avoir lieu au préalable pour constituer un consensus favorable à la suppression. Les administrateurs sont généralement chargés de clore ces discussions, mais des contributeurs non administrateurs en

	règle peuvent le faire sous certaines conditions. Des contributeurs peuvent toutefois demander la suppression d'une page s'ils pensent que celle-ci sera acceptée sans controverse. Dans certains cas, il est possible de supprimer rapidement une page si elle remplit les critères définis par la communauté, qui incluent notamment les pages créées aux seules fins de dénigrer, menacer, intimider ou harceler leur sujet ou une autre entité (Wikipédia, 2020). La Fondation Wikimedia déclare intervenir rarement dans les décisions de la communauté concernant les règles et leur application. Cependant, si la communauté demande une intervention ou la prise en charge d'un utilisateur particulièrement problématique parce qu'il crée des troubles importants ou se livre à un comportement dangereux, elle peut mener une enquête sur l'utilisation des services par l'utilisateur (a) pour déterminer si une infraction aux règles ou à la législation a eu lieu, ou (b) pour se conformer à la législation en vigueur, à une procédure judiciaire ou à une demande des autorités. Des sanctions peuvent être appliquées après l'enquête (voir la section 6 ci-dessous).
4.1 Notifications des suppressions ou des autres décisions de sanction	Sans objet.
4.2 Mécanismes de recours en cas de suppression ou d'autres décisions de sanction	Sans objet.
5. Moyens de détecter les contenus terroristes et extrémistes violents (par exemple, algorithmes de surveillance, contenu généré par les utilisateurs, évaluateurs humains [personnel],	Il appartient à la communauté Wikipédia de procéder au contrôle éditorial, et donc à la détection des contenus contraires aux règles de Wikipédia. Les lecteurs (utilisateurs de Wikipédia qui ne font pas de contribution) peuvent contacter l'équipe bénévole de réponse pour signaler un problème sur un contenu en ligne. La Fondation Wikimedia déclare qu'elle n'assume aucune fonction éditoriale sur ses projets, y compris Wikipédia. Cela signifie qu'« en général », elle ne surveille pas et ne modifie pas le contenu des sites internet de ses projets (Wikimedia Foundation, 2019). Les modérateurs de la communauté Wikipédia n'engendrent aucun coût pour la Fondation Wikimedia. Wikipédia n'est pas membre du GIFCT et ne participe pas au consortium de partage d'empreintes numériques du forum, le Hash Sharing Consortium.

bases de données de partage d'empreintes numériques ou d'adresses URL)	
6. Sanctions ou conséquences en cas de violation des conditions d'utilisation ou des règles de la communauté	La communauté Wikipédia peut envoyer un avertissement, mener une enquête, supprimer des pages créées par des utilisateurs et bloquer ou exclure des utilisateurs qui enfreignent les règles de la communauté. La Fondation Wikimedia peut refuser, désactiver ou restreindre l'accès à la contribution d'un utilisateur qui enfreint ses conditions d'utilisation, interdire à un utilisateur de contribuer à une page ou de modifier un contenu, ou bloquer le compte ou l'accès d'un utilisateur qui a enfreint ses conditions d'utilisation, et prendre des mesures judiciaires à son encontre (y compris envoyer un signalement aux autorités répressives).
7. Publication par le service de rapports de transparence sur les contenus terroristes et extrémistes violents	Non. La Fondation Wikimedia Foundation publie des rapports de transparence (Wikimedia Foundation, n.d.) traitant notamment des demandes de données des utilisateurs et des demandes de modification et de retrait de contenus, mais ils ne mentionnent pas précisément les contenus terroristes et extrémistes violents.
8. Informations ou types de données figurant dans les rapports de transparence	Dans la partie du rapport relative aux demandes de données des utilisateurs, à la section consacrée aux divulgations d'urgence (Emergency disclosures), la Fondation Wikimedia publie le nombre de communications de données des utilisateurs liées à des menaces terroristes. La Fondation Wikimedia contacte proactivement les autorités répressives dès lors qu'elle a connaissance que des projets Wikimedia comportent des informations inquiétantes, telles que des menaces d'explosion. Le nombre indiqué ne correspond toutefois pas au nombre de retraits de contenus terroristes et extrémistes violents.
9. Méthodologies appliquées pour déterminer, calculer ou estimer les informations et données figurant dans	Sans objet.

les rapports de transparence	
10. Fréquence de publication des rapports de transparence	Sans objet.
11. Utilisation du service pour publier des contenus terroristes et extrémistes violents	Information inconnue.

Annexe C – Glossaire

Les définitions et explications ci-après ont pour objet de clarifier certains termes utilisés couramment dans les rapports de transparence sur les contenus terroristes et extrémistes violents.

Blocage – Voir Mesures prises à l'égard de contenus.

Contenus générés par les utilisateurs – Contenus créés, chargés ou partagés par les utilisateurs d'un service de partage de contenus.

Contenu terroriste et extrémiste violent – Dans l'optique du Cadre relatif à l'établissement de rapports de transparence volontaires, tout type d'information numérique servant à la diffusion de contenus terroristes et extrémistes violents, qu'il s'agisse de textes, de vidéos, de contenus audio ou de photos. Il n'existe pas de définition universellement admise du terrorisme ou de l'extrémisme violent, ni, par extension, des contenus terroristes et extrémistes violents. Les entreprises ont toutefois à leur disposition un certain nombre de ressources susceptibles de les aider à choisir et expliciter les définitions du terrorisme et de l'extrémisme violent qu'elles utilisent. Tel est le cas par exemple du Rapport du Rapporteur spécial sur la lutte antiterroriste de 2010 au Conseil des droits de l'homme : Dix pratiques optimales en matière de lutte antiterroriste (Section III. F, « Définition du terrorisme ») ; du document n° 7 du Global Research Network on Terrorism and Technology, *Terrorist Definitions and Designations Lists: What Technology Companies Need to Know* ; et des *Recommandations de Zurich-Londres sur la prévention et la lutte contre l'extrémisme violent et le terrorisme en ligne*, du Forum mondial de lutte contre le terrorisme.

Déclassement – Voir Mesures prises à l'égard de contenus.

Demandes juridiques des pouvoirs publics – Voir Détection et modération, Détection, Réactive.

Démonétisation – Voir Mesures prises à l'égard de contenus.

Déréférencement – Voir Mesures prises à l'égard de contenus.

Désactivation – Voir Mesures prises à l'égard de comptes.

Désactivation de compte – Voir Mesures prises à l'égard de comptes.

Désactivation de contenu – Voir Mesures prises à l'égard de contenus.

Détection et modération – La détection et la modération peuvent intervenir à différents stades et prendre diverses formes. Elles peuvent être quasi simultanées (en cas de recours à des systèmes automatisés, par exemple) ou s'échelonner dans le temps de manière séquentielle (en cas d'examen manuel d'un contenu signalé par un utilisateur). Quelques formes et définitions courantes de détection et de modération sont exposées ci-après.

Détection – On entend par détection le processus d'identification de contenus terroristes et extrémistes violents ou d'activités en ligne connexes sur un service de partage de contenus. Cette détection peut être :

1. Proactive – On parle de détection proactive lorsque les contenus terroristes et extrémistes violents ou les activités en ligne liées à ce type de contenus sont repérés à la faveur d'une détection de routine effectuée par l'entreprise. Elle peut être réalisée grâce à des dispositifs manuels, des outils ou des systèmes hybrides d'examen mis en place par le service de partage de contenus. La détection proactive peut se faire :
    a. Au chargement – La détection proactive au chargement intervient dès qu'un utilisateur tente d'ajouter des contenus terroristes et extrémistes violents ou d'effectuer des actions spécifiques liées à ce type de contenus sur un service en ligne, **avant** qu'ils ne soient partagés avec d'autres utilisateurs ou ne deviennent accessibles. Ce type de détection est essentiellement réalisé à l'aide d'outils automatisés. Une fois que les contenus ou activités ont été détectés, diverses mesures de modération peuvent être prises. Par exemple, si les contenus n'enfreignent pas de manière flagrante ou manifeste les règles de l'entreprise, ils peuvent être aiguillés vers un examen manuel.
    b. Après le chargement – La détection proactive après chargement intervient une fois que les contenus terroristes et extrémistes violents ont été ajoutés sur un service de partage de contenus. Selon les circonstances, cette détection peut avoir lieu **avant** que ces contenus soient partagés avec d'autres utilisateurs ou deviennent accessibles, **ou après**. Là encore, dès lors que les contenus concernés ont été détectés, diverses mesures de modération peuvent être prises.

2. Réactive – On parle de détection réactive lorsque les contenus terroristes et extrémistes violents ou les activités liées à ce type de contenus sont signalés par un tiers au service de partage de contenus. Ce signalement peut émaner d'utilisateurs (voir ci-après les rapports des communautés en ligne) ou d'autres intervenants, comme les organisations de la société civile, les pouvoirs publics, les organismes chargés de faire respecter l'application des lois, les signaleurs de confiance, les organismes de réglementation, les organisations professionnelles, etc. Les rapports émanant d'institutions ou d'organismes publics peuvent prendre la forme de signalements ou de demandes juridiques. S'il n'existe pas toujours de distinction claire entre les deux catégories, la plupart des signalements ou des demandes juridiques correspondent aux paramètres énoncés au titre des deux premiers éléments ci-dessous. Il arrive également que les services de partage de contenus disposent de circuits de signalement ou de canaux d'escalade spéciaux pour des individus, des entités, des types de demandes, des contenus terroristes et extrémistes violents, des activités ou des situations connexes spécifiques – par exemple en cas d'événement terroriste ou extrémiste violent réel ayant des incidences directes en ligne. Dans la mesure où ils dépendent de la façon dont les entreprises conçoivent leurs procédures de signalement, les circuits ou canaux décrits ci-après peuvent légèrement différer ou se recouper.
    a. Demandes juridiques des pouvoirs publics – Les demandes juridiques des pouvoirs publics donnent à un service de partage de contenus l'instruction de retirer des contenus terroristes ou extrémistes violents ou des activités en ligne liées à ce type de contenus qui enfreignent la loi dans une juridiction nationale ou régionale. Ces demandes peuvent prendre diverses formes – dont des avis et des ordonnances – et être fondées sur différents types de législations et de

systèmes juridiques. Elles peuvent émaner d'organismes publics, tels que des institutions publiques, des organismes de réglementation ou d'autres organes administratifs, d'organismes chargés de faire respecter l'application des lois ou de tribunaux nationaux.

b. Signalements des pouvoirs publics – Les signalements des pouvoirs publics sont des demandes par lesquelles une institution ou une autorité publique enjoignent à un service de partage de contenus d'examiner des contenus terroristes ou extrémistes violents ou des activités en ligne liées à ce type de contenus au motif qu'ils pourraient contrevenir aux règles collectives, conditions d'utilisation ou autres documents d'orientation pertinents de l'entreprise. Il peut arriver que les contenus ou activités visés enfreignent également la législation locale.
c. Unités de référence internet – Les unités de référence internet (en anglais *Internet Referral Units*, ou IRU) sont des autorités publiques spécialisées relevant généralement des organismes de contrôle de l'application des lois, chargées d'adresser des signalements aux services de partage de contenus. Ces unités agissent dans les limites de leur mandat et signalent les contenus terroristes et extrémistes violents ou les activités en ligne liées à ce type de contenus qui enfreignent la législation antiterroriste d'un pays donné, mais pour lesquels il est demandé au service concerné de procéder à un examen au regard de ses conditions d'utilisation.
d. Rapports des communautés en ligne – Les rapports ou signalements des communautés sont un mécanisme couramment employé par lequel les utilisateurs signalent des contenus terroristes et extrémistes violents ou des activités en ligne liées à ce type de contenus à un service de partage de contenus.
e. Événement terroriste ou extrémiste violent réel ayant des incidences directes en ligne – Incident à caractère terroriste ou extrémiste violent survenant dans le monde réel et ayant une manifestation simultanée dans l'environnement numérique. Des contenus terroristes ou extrémistes violents sont alors produits par l'auteur ou un complice dans le but de dépeindre un meurtre (ou une tentative de meurtre), un acte de torture ou un préjudice physique grave au nom d'une idéologie, apparaissent comme ayant été conçus, produits et diffusés en vue d'une propagation virale – ou sont devenus viraux –, sont partagés sur l'internet d'une manière qui présage un impact inhabituellement élevé (en termes d'échelle géographique ou de diffusion interplateformes), sont susceptibles de causer un préjudice important à des communautés, et exigent par conséquent une réponse rapide, coordonnée et résolue des acteurs du secteur et des organismes publics compétents. Par exemple, la diffusion en direct de l'attaque de Christchurch a été considérée comme un événement terroriste ou extrémiste violent survenu dans le monde réel, avec des incidences directes en ligne, et appelant une réponse et une action rapides du secteur et des organismes publics compétents.
f. Signaleurs de confiance – Certains services de partage de contenus désignent des signaleurs ou des partenaires qu'ils

jugent suffisamment dignes de confiance, efficaces ou experts d'un type de violation ou de préjudice particulier pour les informer de la présence d'un contenu terroriste ou extrémiste violent ou d'une activité en ligne liée à ce type de contenu qui enfreindrait leurs règles. Le statut de « signaleur de confiance » peut être assorti de privilèges spéciaux, tels que le traitement prioritaire des signalements qu'ils émettent, des fonctions de signalement plus perfectionnées et un dialogue étroit avec le service de partage de contenus au sujet des décisions de modération. Selon le service concerné, les signaleurs de confiance peuvent être des individus, des organisations et/ou des institutions publiques.

3. Détection manuelle – On parle de détection manuelle (également dénommée détection humaine) lorsque des individus repèrent manuellement des contenus terroristes et extrémistes violents ou des activités en ligne liées à ce type de contenus en se fondant sur les règles définies par un service de partage de contenus et sur les ressources et processus internes pertinents, y compris en termes de contrôle qualité. Selon les circonstances, ces personnes peuvent être employées, sous contrat ou désignées pour remplir cette mission.

4. Détection automatisée – On parle de détection automatisée lorsque l'on recourt à des outils technologiques de manière automatique et répétitive, sans déclenchement humain, pour repérer, faire remonter, trier et/ou traiter des contenus terroristes et extrémistes violents ou des activités en ligne liées à ce type de contenus qui enfreignent les règles d'un service de partage de contenus.

Modération – Processus d'examen/d'évaluation des contenus terroristes et extrémistes violents ou des activités en ligne liées à ce type de contenus, puis de prise de décision quant à la suite à y donner à la lumière des règles d'un service de partage de contenus. Les processus de modération et d'examen manuel peuvent être déclenchés par des processus internes d'investigation, par des contrôles de routine, ou à partir d'un système de tri automatisé. Ils peuvent également être déclenchés par une entité tierce externe avertissant ou informant une entreprise de la présence de contenus ou d'activités de ce type susceptibles d'enfreindre ses règles.

1. Modération interne – On parle de modération interne lorsque des contenus terroristes et extrémistes violents ou des activités en ligne liées à ce type de contenus sont examinés/évalués par des équipes de modération ou des administrateurs internes, ou par des organismes ou des services de modération externes, travaillant sous contrat ou sur les instructions d'un service de partage de contenus pour décider des modalités d'application des règles qu'il a définies.

2. Modération par les utilisateurs – On parle de modération par les utilisateurs, ou fondée sur la communauté, lorsque les utilisateurs ou la communauté d'un service de partage de contenus assurent la modération des contenus terroristes et extrémistes violents ou des activités en ligne liées à ce type de contenus directement sur la plateforme du service. Cette modération peut se faire par le biais d'un système de retrait ou de vote qui permet aux utilisateurs d'enregistrer leur approbation ou leur désapprobation.

3. Modération automatisée – On parle de modération automatisée lorsque des outils technologiques sont utilisés de manière automatique et répétitive pour déclencher des mesures face à des contenus terroristes et extrémistes violents ou des activités en ligne liées à ce type de contenus identifiés comme enfreignant les règles du service concerné.

4. Modération manuelle – On parle de modération manuelle (également dénommée modération humaine) lorsque des individus examinent/évaluent manuellement des contenus terroristes et extrémistes violents ou des activités en ligne liées à des contenus de cet ordre générés par des utilisateurs, en se fondant sur les règles du service concerné, sur les ressources et processus internes pertinents et, dans certains cas, sur l'expertise ou la compréhension sociolinguistique du modérateur. Selon les circonstances, ces personnes peuvent être employées, sous contrat ou désignées pour remplir cette mission.

5. Système de modération hybride – Système hybride mêlant la détection et la modération automatisées et manuelles. Il s'agit là de la configuration la plus couramment utilisée par les services de partage de contenus.

6. Modération fondée sur l'activité – On parle de modération fondée sur l'activité lorsque les décisions de modération sont prises à la lumière des activités des utilisateurs liées à des contenus terroristes et extrémistes violents plutôt qu'à partir d'éléments de contenus spécifiques qu'un utilisateur partage. Cela signifie en substance que des mesures peuvent être prises à l'égard de contenus partagés par des utilisateurs et/ou sur des comptes d'utilisateurs bien qu'aucun élément de contenu spécifique n'ait violé à proprement parler la politique du service. Ce type de modération peut s'appuyer sur des méthodes telles que les typologies d'utilisateurs, les comptes ou les signaux d'accès et le profilage de l'environnement.

Diffusion en direct – Utilisation d'un service de partage de contenus pour capter et diffuser le contenu audiovisuel d'un événement en temps réel. Le contenu ainsi transmis est un « direct en streaming ».

Empreinte numérique – Une empreinte numérique est un identifiant unique, souvent assimilé à une signature ou une empreinte digitale, pouvant être créé à partir d'une image ou d'une vidéo numérique.

Événement terroriste ou extrémiste violent réel ayant des incidences directes en ligne – Voir Détection et modération, Détection, Réactive.

Exposé des motifs – Un service de partage de contenus peut fournir un exposé des motifs (de type infraction ou non-infraction aux règles de l'entreprise) à l'utilisateur ayant signalé un contenu, demandé un examen ou publié le contenu, ainsi qu'à tout autre utilisateur concerné et/ou à l'ensemble de la communauté.

Interdiction – Voir Mesures prises à l'égard de comptes.

Masquage – Voir Mesures prises à l'égard de contenus.

Mesures prises à l'égard de comptes – Dans le cadre du traitement des questions liées aux contenus terroristes et extrémistes violents, un service de partage de contenus peut être amené à prendre des mesures à l'égard de l'activité en ligne d'un utilisateur ou d'un compte. Il peut à ce titre appuyer ou récompenser le comportement positif d'un utilisateur, qui aurait par exemple

utilement signalé ou rapporté un contenu problématique. À l'inverse, il peut prendre des mesures pour prévenir ou gérer un comportement négatif, tel que le partage de contenus terroristes et extrémistes violents enfreignant ses règles. Dans ce cas, les mesures peuvent être de plusieurs ordres :

Avertissement – Un message ou une note d'avertissement peut être émis(e) à l'égard d'un compte ayant enfreint les règles d'un service.

Désactivation/mise hors service/suspension – La désactivation – qui peut couvrir le retrait, la suppression, la mise hors service ou la suspension d'un compte – entraîne la clôture effective du compte ayant enfreint les règles. Elle peut être temporaire ou définitive, éligible à des mécanismes de recours ou assortie d'un délai spécifique. Elle peut ou non influer sur l'accessibilité des contributions passées du titulaire du compte sur le service de partage de contenus et être soumise à une obligation de conservation des données à des fins d'application de la loi ou pour des motifs similaires.

Interdiction – L'interdiction a pour effet d'empêcher un utilisateur de se connecter à un service de partage de contenus et/ou de créer et d'utiliser un nouveau compte.

Restriction des privilèges utilisateur – Tout en laissant un compte opérationnel, il est possible de limiter, réduire, suspendre ou retirer certains privilèges qui lui sont associés. Ces privilèges peuvent avoir trait à la possibilité de diffuser des contenus en streaming, d'ajouter des commentaires ou de publier des contenus.

Signalement à un service chargé de l'application des lois – Un utilisateur ou un compte peut faire l'objet d'un signalement à un service chargé de faire respecter l'application des lois en cas d'activité illégale ou de risque imminent pour la sécurité.

Mesures prises à l'égard de contenus – Une fois l'issue du processus de modération connue, le contenu peut soit demeurer dans son état initial sur la plateforme en ligne, soit faire l'objet de mesures prises par le modérateur (personnel interne, entreprise technologique tierce et/ou tiers désigné). Une action peut également être entreprise à titre provisoire en attendant l'issue du processus de modération. Diverses mesures peuvent être mises en place à l'encontre du contenu, parmi lesquelles :

Blocage/désactivation – Il s'agit de limiter ou d'empêcher l'accès à un contenu spécifique par un utilisateur ou un groupe d'utilisateurs. Le géoblocage, par exemple, restreint l'accès des utilisateurs dont les adresses IP sont enregistrées dans une zone géographique donnée. Le contenu peut rester accessible à certains utilisateurs dans des conditions particulières.

Déclassement – En cas de déclassement, le contenu reste disponible sur le service de partage mais bénéficie d'une visibilité moindre. On parle également de déhiérarchisation ou de limitation de visibilité.

Démonétisation – La démonétisation de contenu consiste à limiter la capacité à tirer parti des fonctions de monétisation d'un service de partage de contenus. On peut à ce titre supprimer la possibilité de faire apparaître des publicités à proximité de contenus ne respectant pas les règles applicables (contenus ou autres).

Déréférencement – Retrait d'un contenu, par un service de partage de contenus ou un utilisateur, des listes de recommandations ou de l'index utilisé par les fonctions d'exploration ou de découverte qui permettent aux utilisateurs de rechercher des contenus sur un service.

Masquage/mise en quarantaine – Les notifications émises avant que le contenu ne devienne accessible sont également appelées notes interstitielles. Le contenu masqué derrière une note interstitielle peut devenir accessible si des conditions particulières sont remplies – par exemple,

si les utilisateurs déclarent leur âge ou reconnaissent que le contenu peut être choquant. Le contenu peut également être placé en quarantaine ou masqué derrière une notification afin d'indiquer qu'il n'est pas accessible aux utilisateurs car il est en cours d'examen ou il enfreint les règles d'une entreprise.

Notification – Un modérateur peut ajouter une notification à des contenus générés par des utilisateurs afin d'informer les autres utilisateurs qu'il peut être sensible, perturbant, faux, inadapté au jeune public, ou problématique du point de vue des attentes de la communauté, bien que n'enfreignant pas les règles de l'entreprise.

Retrait – Le retrait désigne le processus par lequel un service de partage de contenus retire un contenu de sorte qu'aucun utilisateur ne puisse y accéder. Le caractère provisoire ou définitif d'un retrait dépend des règles et des mécanismes de recours du service de partage de contenus, ainsi que de la légalité du contenu.

Mise en quarantaine – Voir Mesures prises à l'égard de contenus.

Modération – Voir Détection et modération, Modération.

Notification – Voir Mesures prises à l'égard de contenus.

Rapports des communautés en ligne – Voir Détection et modération, Détection, Réactive.

Recours et réexamen – Processus par lequel un ou plusieurs utilisateur(s) jugeant l'issue d'une décision de modération incorrecte peu(ven)t en solliciter le réexamen. Les services de partage de contenus offrant des possibilités de recours ou de réexamen des décisions peuvent procéder à une révision automatisée et/ou humaine. Celle-ci peut être conduite en interne, par le service et/ou dans le cadre d'un dispositif adapté impliquant les membres de la communauté d'utilisateurs, ou par un organisme externe indépendant, y compris par les autorités judiciaires dans les pays concernés. Si, à l'issue du réexamen, la décision est prise d'annuler, d'invalider ou de modifier le résultat initial du processus de modération, des formes de règlement ou de réparation peuvent se mettre en place, telles le rétablissement du contenu ou du compte, ou d'autres mesures prises à l'égard des contenus ou des comptes (voir plus haut).

Règles de l'entreprise – Les règles de l'entreprise sont également dénommées standards, règles collectives, politique d'utilisation acceptable, conditions générales d'utilisation ou modalités d'utilisation. Ces règles servent généralement à énoncer les attentes quant aux contenus ou activités qu'il est autorisé ou non de publier ou de mener à bien en rapport avec un service ou un produit d'une entreprise. Elles peuvent également préciser les mesures que l'entreprise peut prendre à l'égard des contenus ou des comptes, ainsi que les mécanismes de notification des utilisateurs et de recours mis à leur disposition.

Rétablissement – Rétablissement de contenus ou de comptes et/ou annulation des mesures prises à leur égard.

Retrait – Voir Mesures prises à l'égard de contenus.

Services de partage de contenus – Les services de partage de contenus désignent tout service en ligne permettant le transfert et la diffusion de contenus, sous quelque forme que ce soit, selon un rapport « un à un », un à une cible réduite, ou « un à *n* ».

Signalements des pouvoirs publics – Voir Détection et modération, Détection, Réactive.

Signaleurs de confiance – Voir Détection et modération, Détection, Réactive.

Suspension – Voir Mesures prises à l'égard de comptes.

Unité de référence internet – Voir Détection et modération, Détection, Réactive.

Références

4chan. (s.d.). *'Advertise - 4chan'*. Consulté le August 31, 2019, sur http://www.4chan.org/advertise

4chan. (s.d.). *Frequently Asked Questions*. Récupéré sur https://www.4channel.org/faq

Ahmed, M. (2020, janvier 31). *After Christchurch: How Policymakers Can Respond to Online Extremism*. Récupéré sur Tony Blair Institute for Global Change: https://institute.global/policy/after-christchurch-how-policymakers-can-respond-online-extremism

Alexa. (2019). *The top 500 sites on the web.* Récupéré sur Alexa: https://www.alexa.com/topsites/global;0

Alexander, J. (2019, août 12). *Verizon is selling Tumblr to WordPress' owner*. Récupéré sur The Verge: https://www.theverge.com/2019/8/12/20802639/tumblr-verizon-sold-wordpress-blogging-yahoo-adult-content

Amazon. (s.d.). *Amazon.com Help: Law Enforcement Information Requests*. Récupéré sur Amazon: https://www.amazon.com/gp/help/customer/display.html?nodeId=GYSDRGWQ2C2CRYEF

Apple. (s.d.). *Privacy - About Apple's Transparency Report*. Récupéré sur Apple: https://www.apple.com/legal/transparency/about.html

Arthur, R. (2019, juillet 10). *We Analyzed More Than 1 Million Comments on 4chan. Hate Speech There Has Spiked by 40% Since 2015.* Récupéré sur Vice: https://www.vice.com/en_us/article/d3nbzy/we-analyzed-more-than-1-million-comments-on-4chan-hate-speech-there-has-spiked-by-40-since-2015

Automattic. (s.d.). *Rapport de transparence*. Récupéré sur Automattic: https://transparency.automattic.com/

Barnes, L. (2019, janvier 17). *One month after controversial adult-content purge, far-right pages are thriving on Tumblr*. Récupéré sur Think Progress: https://thinkprogress.org/far-right-content-survived-tumblr-purge-36635e6aba4b/

Barret, P. M. (2020). *Regulating Social Media: The Fight overn Section 230 - and Beyond.* NYU / STERN - Center for Business and Human Rights.

Bennett, C. e. (2019). *Extremism, George Washington University.* Récupéré sur https://extremism.gwu.edu/sites/g/files/zaxdzs2191/f/EncryptedExtremism.pdf

Bicknell, Z. (2018, septembre 27). *What Video Platform Should I Use?* Récupéré sur The UK Domain: https://www.theukdomain.uk/what-video-platform-should-i-use/

British Broadcasting Corporation (BBC). (2019, octobre 10). *Germany shooting: 2,200 people watched on Twitch*. Récupéré sur BBC: https://www.bbc.com/news/technology-49998284

Carmen, A. (2015, décembre 9). *Filtered extremism: how ISIS supporters use Instagram*. Récupéré sur The Verge: https://www.theverge.com/2015/12/9/9879308/isis-instagram-islamic-state-social-media

Cheah, M. (2019, juin 26). *Important updates to our content guidelines - Vimeo Blog*. Récupéré sur Vimeo: https://vimeo.com/blog/post/important-updates-to-our-content-guidelines/

Chen, W. (2020, avril 1). *The top Chinese short-video apps in 2020 vying to grab your attention with fast content*. Récupéré sur KrASIA: https://kr-asia.com/the-top-chinese-short-video-apps-in-2020-vying-to-grab-your-attention-with-fast-content

Christchurch Call. (2019). *Christchurch Call.* Récupéré sur https://www.christchurchcall.com/call.html

Clicky, S. (2017, décembre 6). *Tackling Extremist Content on WordPress.com*. Récupéré sur Rapport de transparence: https://transparency.automattic.com/2017/12/06/tackling-extremist-content-on-wordpress-com/

Clifford, B., & Powell, H. (2019). *Encrypted Extremism - Inside the English-Speaking Islamic State Ecosystem on Telegram.* The George Washington University, Program on Extremism.

Commission européenne. (2020). *Proposition de Règlement du Parlement européen et du Conseil relatif à un marché intérieur des services numériques (Législation sur les services numériques) et modifiant la directive 2000/31/CE.*

Commission européenne. (2020). *Rapport de la Commission au Parlement européen et au Conseil présenté conformément à l'article 29, paragraphe 1, de la directive (UE) 2017/541 du Parlement européen et du Conseil du 15 mars 2017 relative à la lutte contre le terrorisme et remplaçant la décision-cadre 2002/475/JAI du Conseil et modifiant la décision 2005/671/JAI du Conseil.*

Commission européenne. (2020, juin 02). *The Digital Services Act package*. Récupéré sur https://ec.europa.eu/digital-single-market/en/digital-services-act-package

Conseil de Sécurité des Nations Unies. (s.d.). *Liste récapitulative du Conseil de sécurité des Nations Unies.* Récupéré sur Conseil de Sécurité des Nations Unies: https://www.un.org/securitycouncil/fr/content/un-sc-consolidated-list

Counter Extremism Project. (2018, août 17). *On Anniversary Of Barcelona Attacks, ISIS Continues Its Expansion*. Récupéré sur Counter Extremism Project: https://www.counterextremism.com/press/anniversary-barcelona-attacks-isis-continues-its-expansion

Counter Terrorism Project. (s.d.). *Extremists & Online Propaganda.* Récupéré sur Counter Terrorism Project: https://www.counterextremism.com/extremists-online-propaganda

Cox, J. (2019, avril 19). *36 Days After Christchurch, Terrorist Attack Videos Are Still on Facebook*. Récupéré sur Vice: https://www.vice.com/en_us/article/43jdbj/christchurch-attack-videos-still-on-facebook-instagram

Cuthbertson, A. (2019, décembre 03). *TikTok secretly loaded with Chinese surveillance software, lawsuit claims*. Récupéré sur Independent: https://www.independent.co.uk/life-style/gadgets-and-tech/news/tiktok-china-data-privacy-lawsuit-bytedance-a9230426.html

Datanyze. (2020, septembre). *Market Share / File Sharing.* Récupéré sur Datanyze: https://www.datanyze.com/market-share/file-sharing--198/Datanyze%20Universe

Daum Kakao. (s.d.). *Transparency Report, Kakao Privacy Policy*. Récupéré sur Kakao: http://privacy.daumkakao.com/en/transparence/report/request

DCMS. (2020, février 12). *Online Harms White Paper - Initial consultation response*. Récupéré sur https://www.gov.uk/government/consultations/online-harms-white-paper/public-feedback/online-harms-white-paper-initial-consultation-response

Dearden, L. (2019, août 9). *Far-right extremists 'encouraged copycat terror attacks' after Christchurch mosque shootings.* Récupéré sur The Independent:

https://www.independent.co.uk/news/uk/crime/far-right-terror-plots-uk-muslims-christchurch-attack-white-a9050511.html

Department of the Prime Minister and Cabinet, A. (2019, juin 21). *Australian Taskforce to Combat Terrorist and Extreme Violent Material Online.* Consulté le June 5, 2019, sur https://www.pmc.gov.au/sites/default/files/publications/combat-terrorism-extreme-violent-material-online.pdf

DeviantArt Media Kit. (s.d.). *There's No Place Like DeviantArt*. Récupéré sur https://deviantartads.com/

DeviantArt. (s.d.). *What happens when my account is banned?* Récupéré sur DeviantArt: https://www.deviantartsupport.com/en/article/what-happens-when-my-account-is-banned

DeviantArt. (s.d.). *What is your policy around account suspensions?* Récupéré sur DeviantArt: https://www.deviantartsupport.com/en/article/what-is-your-policy-around-account-suspensions

DeviantArt. (s.d.). *What policy guidelines are there on comments, Journals, statuses, and general interactions?* Récupéré sur DeviantArt: https://www.deviantartsupport.com/en/article/what-policy-guidelines-are-there-on-comments-journals-statuses-and-general-interactions

Dilger, D. E. (2015, novembre 21). *Another security manual recommends using Apple iMessage: this time, ISIS*. Récupéré sur appleinsider: https://appleinsider.com/articles/15/11/21/another-security-manual-recommends-using-apple-imessage-this-time-isis-

Discord. (2019). *Discord Transparency Report: Jan 1 — April 1*. Récupéré sur Discord Blog: https://blog.discordapp.com/discord-transparency-report-jan-1-april-1-4f288bf952c9?gi=e7efc9d05321

Discord. (2020). *Discord Transparency Report: April — Dec 2019*. Récupéré sur https://blog.discord.com/discord-transparency-report-april-dec-2019-7e6d43a9bcb8

Dropbox. (s.d.). *Transparency Overview*. Récupéré sur Dropbox: https://www.dropbox.com/en_GB/transparency

Dropbox. (s.d.). *Who can see the stuff in my Dropbox account? Dropbox Help*. Récupéré sur Dropbox: https://help.dropbox.com/accounts-billing/security/file-access

Duarte, N., Llanso, E., & Loup, A. (2017). *Mixed Messages? The Limits of Automated Social Media Cotent Analysis.* Center for Democracy & Technology.

EDRi. (2019, octobre 17). *Trilogues on terrorist content: Upload or re-upload filters? Eachy peachy*. Récupéré sur EDRi: https://edri.org/our-work/trilogues-on-terrorist-content-upload-or-re-upload-filters-eachy-peachy/

Electronic Frontier Foundation. (2020, octobre). *Urgent: EARN IT Act Introduced in House of Representatives*. Récupéré sur https://www.eff.org/deeplinks/2020/10/urgent-earn-it-act-introduced-house-representatives

Elmer-Dewitt, P. (2019, janvier 17). *Information: Facebook's Messenger has overtaken Apple's iMessage.* Récupéré sur 247wallstreet.com: https://247wallst.com/technology-3/2019/01/17/apple-facebook-messaging/

Facebook. (2017-2020). *Community Standards Enforcement Report - Dangerous Organisations: Terrorism and Organised Hate.* Récupéré sur Facebook: https://transparency.facebook.com/community-standards-enforcement#terrorist-propaganda

Facebook. (2018, novembre 8). *Hard Questions: What Are We Doing to Stay Ahead of Terrorists?* Récupéré sur Facebook: https://about.fb.com/news/2018/11/staying-ahead-of-terrorists/

Facebook. (2019, septembre 17). *Combating Hate and Extremism*. Récupéré sur Facebook: https://about.fb.com/news/2019/09/combating-hate-and-extremism/

Facebook. (2019, mai 23). *Measuring Prevalence of Violating Content on Facebook*. Récupéré sur Facebook: https://about.fb.com/news/2019/05/measuring-prevalence/

Facebook. (2020, mai 12). *An Update on Combating Hate and Dangerous Organizations*. Récupéré sur Facebook: https://about.fb.com/news/2020/05/combating-hate-and-dangerous-organizations/

Facebook. (2020). *Community Standards Enforcement Report: Dangerous Organizations*. Consulté le June 5, 2020, sur https://transparency.fb.com/fr-fr/policies/community-standards/dangerous-individuals-organizations/

Facebook. (s.d.). *Standards de la communauté 2. Individus et organismes dangereux.* Récupéré sur Site web de Facebook: https://www.facebook.com/communitystandards/dangerous_individuals_organizations

Facebook. (s.d.). *Standards de la communauté, 1. Violence et provocation.* Récupéré sur Site web de Facebook: https://www.facebook.com/communitystandards/credible_violence

Facebook. (s.d.). *Understanding the Community Standards Enforcement Report.* Récupéré sur Facebook Transparency: https://transparency.facebook.com/community-standards-enforcement/guide

Fisher-Birch, J. (2018, mars 13). *Terror on Tumblr*. Récupéré sur Counter Terrorism Project: https://www.counterextremism.com/blog/terror-tumblr

Frier, S. (2018, avril 4). *Facebook Scans the Photos and Links You Send on Messenger.* Récupéré sur Bloomberg: https://www.bloomberg.com/news/articles/2018-04-04/facebook-scans-what-you-send-to-other-people-on-messenger-app

G20. (2017). *The Hamburg G20 Leaders' Statement on Countering Terrorism*. Récupéré sur https://www.mofa.go.jp/files/000271330.pdf

G20. (2019). *G20 Osaka Leaders' Statement on Preventing Exploitation of the Internet for Terrorism and Violent Extremism Conducive to Terrorism (VECT)*. Récupéré sur Digital Watch Observatory: https://dig.watch/instruments/g20-osaka-leaders-statement-preventing-exploitation-internet-terrorism-and-violent

G7. (2019). *Résumé de la Présidence des ministres du numérique du G7*. Récupéré sur https://www.economie.gouv.fr/files/files/2019/G7/G7Num/Chairs_summary_version_finale_FR.pdf

GIFCT. (2020). *GIFCT Transparency Report - July 2020*. Récupéré sur https://gifct.org/transparency/

GIFCT. (2021). *Membership*. Récupéré sur https://gifct.org/membership/

GIFCT. (s.d.). *Global Internet Forum to Counter Terrorism: Evolving an Institution*. Récupéré sur https://gifct.org/about/

GIFCT. (s.d.). *Join Tech Innovation*. Récupéré sur https://gifct.org/joint-tech-innovation/

Google. (2010-2020, Janvier-mai). *Demandes gouvernementales de suppression de contenu - Transparence des informations Google*. Récupéré sur Google: https://transparencyreport.google.com/government-removals/overview?hl=fr_FR

Google. (2019, janvier 22). *Conditions d'utilisation supplémentaires de Google Drive*.

Récupéré sur Google: https://www.google.com/intl/fr/drive/terms-of-service/

Google. (s.d.). *Aide Éditeurs Docs - Demander l'examen d'un rapport pour infraction au règlement*. Récupéré sur Google: https://support.google.com/docs/answer/2463328?hl=fr&ref_topic=1360897

Google. (s.d.). *Aide Éditeurs Docs - Règlement du programme concernant l'utilisation abusive et mise en application*. Récupéré sur Google: https://support.google.com/docs/answer/148505?visit_id=637064013896463652-1393240150&rd=1

Google. (s.d.). *Aide Éditeurs Docs - Signaler une infraction*. Récupéré sur Google: https://support.google.com/docs/answer/2463296?hl=fr&ref_topic=1360897

Google. (s.d.). *Transparence des informations*. Récupéré sur https://transparencyreport.google.com/?hl=fr_FR

Google. (s.d.). *Transparence des informations*. Récupéré sur https://transparencyreport.google.com/?hl=fr_FR

Google, Youtube. (2017-2020). *Google Transparency Report - Flags*. Récupéré sur Google transparency Report: https://transparencyreport.google.com/youtube-policy/flags?request_examples=year:;flagging_reason:7;flagger_type:&lu=request_examples

Google, Youtube. (2019, juin 5). *Our ongoing work to tackle hate*. Récupéré sur YouTube blog: https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate

Google, Youtube. (2020). *Activer ou désactiver le mode restreint.* Récupéré sur Google, Youtube: https://support.google.com/youtube/answer/174084?hl=fr

Google, Youtube. (2020). *Faire appel des actions pour non-respect du règlement de la communauté.* Récupéré sur Google, Youtube: https://support.google.com/youtube/answer/185111?hl=fr

Google, Youtube. (2020). *Signaler un contenu inapproprié.* Récupéré sur Google, Youtube: https://support.google.com/youtube/answer/2802027?hl=fr

Google, Youtube. (2020). *YouTube Trusted Flagger program.* Récupéré sur Google, Youtube: https://support.google.com/youtube/answer/7554338?&ref_topic=2803138

Google, YouTube. (s.d.). *Aide YouTube - Fonctionnalités limitées pour certaines vidéos*. Récupéré sur Google, YouTube: https://support.google.com/youtube/answer/7458465?hl=fr

Google, YouTube. (s.d.). *Community Guidelines strike basics - YouTube Help*. Récupéré sur Google, YouTube: https://support.google.com/youtube/answer/2802032

Google, Youtube. (s.d.). *Sélection de règles : Incitation à la haine*. Récupéré sur Google: https://transparencyreport.google.com/youtube-policy/featured-policies/hate-speech?hl=fr_FR

Google, YouTube. (s.d.). *Sélection de règles : Contenu extrémiste violent*. Récupéré sur Google, YouTube: https://transparencyreport.google.com/youtube-policy/featured-policies/violent-extremism?hl=fr_FR&policy_removals=period:Y2019Q2&lu=policy_removals

Google/ YouTube. (2020). *Règles concernant les organisations criminelles violentes.* Récupéré sur Google/Aide YouTube: https://support.google.com/youtube/answer/9229472?hl=fr&ref_topic=9282436

Gouvernement de l'Australie, F. R. (2019). *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019.* Récupéré sur

https://www.legislation.gov.au/Details/C2019A00038

Gouvernement du Canada. (2021). *Réglementation des plateformes de médias sociaux*. Récupéré sur https://rechercher.ouvert.canada.ca/fr/qp/id/pch,PCH-2020-QP-00084

Gouvernement du Royaume-Uni. (2019, avril). *Online Harms White Paper.* Consulté le June 4, 2019, sur assets.publishing.service.gov.uk: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf

Grüll, P. (2020, juin 19). *German online hate speech reform criticised for allowing 'backdoor' data collection*. Récupéré sur Euractiv: https://www.euractiv.com/section/data-protection/news/german-online-hate-speech-reform-criticised-for-allowing-backdoor-data-collection/

Harwell, D., & Romm, T. (2019, septembre 15). *TikTok's Beijing roots fuel censorship suspicion as it builds a huge U.S. audience*. Récupéré sur The Washington Post: https://www.washingtonpost.com/technology/2019/09/15/tiktoks-beijing-roots-fuel-censorship-suspicion-it-builds-huge-us-audience/

Hatmaker, T. (2019). *This led to Reddit administrators banning the entire community in question from the site.* Récupéré sur The Tech Crunch: https://techcrunch.com/2019/03/15/reddit-watchpeopledie-subreddit-gore/

Hayden, M. E. (2019, juin 27). *Far-Right Extremists Are Calling for Terrorism on the Messaging App Telegram*. Récupéré sur Southern Poverty Law Center: https://www.splcenter.org/hatewatch/2019/06/27/far-right-extremists-are-calling-terrorism-messaging-app-telegram

Hayden, M. E. (2019, mai 21). *Mysterious Neo-Nazi Advocated Terrorism for Six Years Before Disappearance*. Récupéré sur Southern Poverty Law Center: https://www.splcenter.org/hatewatch/2019/05/21/mysterious-neo-nazi-advocated-terrorism-six-years-disappearance

Hern, A. (2019, septembre 25). *Revealed: how TikTok censors videos that do not please Beijing*. Récupéré sur The Guardian: https://www.theguardian.com/technology/2019/sep/25/revealed-how-tiktok-censors-videos-that-do-not-please-beijing

Huang, F. (2018, novembre 27). *China's Most Popular App Is Full of Hate*. Récupéré sur Foreign Policy: https://foreignpolicy.com/2018/11/27/chinas-most-popular-app-is-full-of-hate/

Hymas, C. (2019, mai 11). *Isil extremists using Instagram to promote jihad and incite support for terror attacks on the West*. Récupéré sur The Telegraph: https://www.telegraph.co.uk/news/2019/05/11/isil-extremists-using-instagram-promote-jihad-incite-support/

Instagram. (2019, juillet 18). *Changes to Our Account Disable Policy*. Récupéré sur Instagram: https://instagram-press.com/blog/2019/07/18/changes-to-our-account-disable-policy/

Iqbal, M. (2020, juillet 23). *Twitch Revenue and Usage Statistics (2020).* Récupéré sur Business of Apps: https://www.businessofapps.com/data/twitch-statistics/

ISD Global. (s.d.). *Powering solutions to extremism and polarisation*. Récupéré sur ISD Global: https://www.isdglobal.org/

Kallas, P. (2020, avril 9). *Top 15 Most Popular Social Networking Sites and Apps [2020] @ Dreamgrow.* Récupéré sur dreamgrow.com: https://www.dreamgrow.com/top-15-

most-popular-social-networking-sites/

Katz, R. (2018, octobre 10). *To Curb Terrorist Propaganda Online, Look to YouTube. No, Really.* Récupéré sur Wired: https://www.wired.com/story/to-curb-terrorist-propaganda-online-look-to-youtube-no-really/

Katz, R. (2019, septembre 1). *A Growing Frontier for Terrorist Groups: Unsuspecting Chat Apps*. Récupéré sur Wired: https://www.wired.com/story/terrorist-groups-prey-on-unsuspecting-chat-apps/

Kemp, S. (2019, janvier 31). *Digital 2019: Global Digitial Overview*. Récupéré sur https://datareportal.com/reports/digital-2019-global-digital-overview

Kemp, S. (2019, janvier 31). *Digital 2019: Q3 Global Digital Statshot.* Récupéré sur datareportal.com: https://datareportal.com/reports/digital-2019-q3-global-digital-statshot

Kemp, S. (2020, juillet 21). *More than Half of the People on Earth now Use Social Media.* Récupéré sur https://datareportal.com/reports/more-than-half-the-world-now-uses-social-media

Kenny, K. (2019, avril 30). *How can upcoming social media efforts be 'global' if they ignore Asia?* Récupéré sur Stuff.co.nz: https://www.stuff.co.nz/national/christchurch-shooting/112284082/how-can-upcoming-social-media-efforts-be-global-if-they-ignore-asia

Kenyon, M. (2020, mai 7). *WeChat Surveillance Explained*. Récupéré sur The Citizen Lab: https://citizenlab.ca/2020/05/wechat-surveillance-explained/

Kinsta. (2011-2019). *Wordpress Market Share Statistics (2011-2019).* Récupéré sur Kinsta: https://kinsta.com/wordpress-market-share/

Kitsune, L. (2017, octobre 11). *New Notifications and Reporting Updates by Lauren Kitsune on DeviantArt*. Récupéré sur https://www.deviantart.com/laurenkitsune/journal/New-Notifications-and-Reporting-Updates-706864447

Knockel, J. L.-N. (2018, août 14). *(Can't) Picture This, An Analysis of Image Filtering on WeChat Moments.* Récupéré sur The Citizen Lab: https://citizenlab.ca/2018/08/cant-picture-this-an-analysis-of-image-filtering-on-wechat-moments/

Knockel, J., Parsons, C., Ruan, L., Xiong, R., Crandall, J., & Deibert, e. R. (2020, mai 7). *We Chat, They Watch How International Users Unwittingly Build up WeChat's Chinese Censorship Apparatus*. Récupéré sur Citizen Lab: https://citizenlab.ca/2020/05/we-chat-they-watch/

Knockell, J. M.-N. (2015). Every Rose Has Its Thorn: Censorship and Surveillance on Social VideoPlatforms in China. Récupéré sur https://www.usenix.org/system/files/conference/foci15/foci15-paper-knockel.pdf

Kock, R. (2020, juillet 23). *TikTok and the privacy perils of China's first international social media platform*. Récupéré sur ProtonMail: https://protonmail.com/blog/tiktok-privacy/?utm_campaign=ww-en-2a-generic-coms_soc-social_organic&utm_content=&utm_medium=soc&utm_source=twitter&utm_term=1595520917

Lange, D. (2017, mai 22). *Quora's Tolerance Of Terror Support*. Récupéré sur Israellycool.com: https://www.israellycool.com/2017/05/22/quoras-tolerance-of-terror-support/

Le Monde. (2021, janvier 18). *Haine en ligne : des obligations de transparence pour les réseaux sociaux*. Récupéré sur

https://www.lemonde.fr/politique/article/2021/01/18/haine-en-ligne-des-obligations-de-transparence-pour-les-reseaux-sociaux_6066656_823448.html

Liao, S. (2018, février 28). *Discord shuts down more neo-Nazi, alt-right servers*. Récupéré sur The Verge: https://www.theverge.com/2018/2/28/17062554/discord-alt-right-neo-nazi-white-supremacy-atomwaffen

LINE. (2019-2020). *LINE Content Moderation Report*. Récupéré sur https://linecorp.com/en/security/moderation/2019h1

LINE. (s.d.). *Help Center*. Récupéré sur Line: https://help.line.me/line/android/categoryId/20000132/3/pc?lang=en

Lix Xan Wong, K., & Shields Dobson, A. (2019). We're just data: Exploring China's social credit system in relation to digital platform ratings cultures in Westernised democracies. *Global Media and China, 4*(2), 220-232.

Lokot, T. (2014, septembre 12). *Vkontakte, a Russian social network, is hosting ISIS accounts that were kicked off of Facebook and Twitter*. (M. J. Rosenthal, Éditeur) Récupéré sur PRI: https://www.pri.org/stories/2014-09-12/isis-internet-army-has-found-safe-haven-russian-social-networks-now

Manileve, V. (2016, 15 juillet). *The Problem With Snapchat's Coverage of the Terror in Nice*. Récupéré sur Slate: https://slate.com/technology/2016/07/did-snapchat-show-its-users-too-much-from-the-tragedy-in-nice.html

Marketing Land. (2018, septembre 18). *Quora Introduces Broad Targeting, Says Audience Hits 300 Million Monthly Users.* Récupéré sur marketingland.com: https://marketingland.com/quora-introduces-broad-targeting-says-audience-hits-300-million-monthly-users-248261

Marketing to China. (2020, mars 21). *Top 10 Chinese Social Media for Marketing (updated 2020)*. Récupéré sur https://www.marketingtochina.com/top-10-social-media-in-china-for-marketing/

Marshall, C. (2019, mai 28). *Twitch suspends streaming for new users as it fights off Artifact trolls*. Récupéré sur Twitch: https://www.polygon.com/2019/5/28/18643198/twitch-artifact-section-stream-suspended

Medium. (2015, janvier 5). *Medium's Transparency Report (2014)*. Récupéré sur Medium: https://medium.com/transparency-report/mediums-transparency-report-438fe06936ff

Medium. (s.d.). *Controversial, Suspect and Extreme Content, Medium Help Center*. Récupéré sur Medium: https://help.medium.com/hc/en-us/articles/360018182453

Meetup. (2017, avril 24). *Introducing Meetup's Inaugural Transparency Report*. Récupéré sur The Meetup Blog: http://blog.meetup.com/inaugural-transparency-report/

Meetup. (2019, juin 1). *Conditions générales de service*. Récupéré sur Meetup: https://help.meetup.com/hc/fr-fr/articles/360027447252-Conditions-g%C3%A9n%C3%A9rales-de-service

Microsoft. (2016, mai 20). *Microsoft's approach to terrorist content online, Microsoft on the Issues*. Récupéré sur Microsoft: https://blogs.microsoft.com/on-the-issues/2016/05/20/microsofts-approach-terrorist-content-online/#sm.000del1ea19zbe4duja1ve96fcc1l

Microsoft. (2019, January to June). *Contents Removals Request Report, Microsoft CSR.* Récupéré sur Microsoft: https://www.microsoft.com/en-us/corporate-responsibility/content-removal-requests-report

Microsoft. (2021). *Digital Safety Content Report*. Récupéré sur https://www.microsoft.com/en-

us/corporate-responsibility/digital-safety-content-report?activetab=pivot_1:primaryr4

Miller, J. (2014, juin 25). *Can Iraqi militants be kept off social media sites?* Récupéré sur BBC News.

OCDE. (2020). Current approaches to terrorist and violent extremist content among the global top 50 online content-sharing services. Dans *Documents de travail de l'OCDE sur l'économie numérique.* Éditions OCDE, Paris. doi:https://dx.doi.org/10.1787/68058b95-en

Odnoklassniki. (s.d.). *Help Centre*. Récupéré sur https://ok.ru/help/54/367

Office des Nations Unies contre la drogue et le crime. (2012). *The use of the Internet for terrorist purposes.* Nations Unies. Vienne: Office des Nations Unies à Vienne.

Patriquin, M. (2021, février 1). *With new legislation, Steven Guilbeault will make few friends in Big Tech*. Récupéré sur https://financialpost.com/technology/with-new-legislation-steven-guilbeault-will-make-few-friends-in-big-tech

Penetrum Security. (s.d.). *Petrum Security Analysis of TikTok versions 10.0.8 -15.2.3.*

Perez, S. (2019, juin 5). *Skype publicly launches screen sharing on iOS and Android*. Récupéré sur Tech Crunch: https://techcrunch.com/2019/06/05/skype-publicly-launches-screen-sharing-on-ios-and-android/?guccounter=1

Perez, S. (2020, mars 11). *TikTok to open a 'Transparency Center' where outside experts can examine its content moderation practices*. Récupéré sur Tech Crunch: https://techcrunch.com/2020/03/11/tiktok-to-open-a-transparency-center-where-outside-experts-can-examine-its-moderation-practices/

Pew Research Center. (2016, janvier 14). *Wikipedia at 15: Millions of readers in scores of languages.* Récupéré sur https://www.pewresearch.org/fact-tank/2016/01/14/wikipedia-at-15/

Pinterest. (2014-2020, January to March). *Transparency Report - Pinterest help*. Récupéré sur Pinterest: https://help.pinterest.com/en-gb/article/transparency-report

Pinterest. (s.d.). *Suspension de compte*. Récupéré sur https://help.pinterest.com/fr/article/account-suspension

Powell, B. C. (2019). *Encrypted Extremism - Inside the English-Speaking Islamic State Ecosystem on Telegram.* The George Washington University.

Quay-de la Vallee, H., & Azarmi, M. (2020, août 25). *The New EARN IT Act Still Threatens Encryption and Child Exploitation Prosecutions*. Récupéré sur https://cdt.org/insights/the-new-earn-it-act-still-threatens-encryption-and-child-exploitation-prosecutions/

Quora. (s.d.). *How does Quora Moderation make decisions about edit-blocks and bans? How does someone appeal this decision?* Récupéré sur Quora: https://www.quora.com/How-does-Quora-Moderation-make-decisions-about-edit-blocks-and-bans-How-does-someone-appeal-this-decision

Reddit. (2019). *Transparency Report 2019*. Récupéré sur Reddit: https://www.redditinc.com/policies/transparency-report-2019

Reddit Inc. (2017, avril 17). *Moderator Guidelines for Healthy Communities*. Récupéré sur Reddit: https://www.redditinc.com/policies/moderator-guidelines

Reddit Inc. (2018, January to December). *Transparency Report 2018*. Récupéré sur Reddit: https://www.redditinc.com/policies/transparency-report-2018

Reddit Inc. (2020). *Transparency Report 2020*. Récupéré sur

https://www.redditinc.com/policies/transparency-report-2020-1

Reddit Inc. (s.d.). *AutoModerator*. Récupéré sur Reddit help: https://mods.reddithelp.com/hc/en-us/articles/360002561632-AutoModerator

Reddit Inc. (s.d.). *Quarantined Subreddits*. Récupéré sur Reddit Help: https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/quarantined-subreddits

Ruan, L. J.-N. (2016). *One App, Two Systems, How WeChat uses one censorship policy in China and another internationally.* Récupéré sur The Citizen Lab: https://citizenlab.ca/2016/11/wechat-china-censorship-one-app-two-systems/

Ruan, L., Knockel, J., Ng, J. Q., & Crete-Nishihata, e. M. (2016). *One App, Two Systems, How WeChat uses one censorship policy in China and another internationally.* Récupéré sur The Citizen Lab: https://citizenlab.ca/2016/11/wechat-china-censorship-one-app-two-systems/

Santa Clara University's High Tech Law Institute. (s.d.). *The Santa Clara Principles On Transparency and Accountability in Content Moderation*. Récupéré sur santaclaraprinciples.org: https://santaclaraprinciples.org/

Singh, M. (2020, avril 24). *Telegram hits 400M monthly active users.* Récupéré sur Tech Crunch: https://techcrunch.com/2020/04/24/telegram-hits-400-million-monthly-active-users/

Site Intelligence Group Enterprise. (2018, décembre 11). *IS-linked Media Group Makes Foray onto Viber Messenger - Dark Web and Cyber Security*. Récupéré sur Site Intelligence Group Enterprise: https://ent.siteintelgroup.com/Dark-Web-and-Cyber-Security/is-linked-media-group-makes-foray-onto-viber-messenger.html

Sky News. (2020, mai 19). *FBI unlocks terrorist's iPhones and finds al Qaeda links - 'no thanks to Apple'*. Récupéré sur Sky News: https://news.sky.com/story/fbi-unlocks-terrorists-iphones-and-finds-al-qaeda-links-no-thanks-to-apple-11990818

Snap Inc. (2015-2020). *Rapport sur la transparence (1 janvier 2020 – 30 juin 2020)*. Récupéré sur Snap Inc.: https://www.snap.com/fr-FR/privacy/transparency

Snap Inc. (s.d.). *Centre de sécurité, Signaler un problème de sécurité*. Récupéré sur Snap Inc.: https://www.snap.com/fr-FR/safety/safety-reporting

Solsman, J. E. (2018, juillet 14). *« Smule May Be the Biggest Music App You Haven't Heard Of ».* Récupéré sur CNET: https://www.cnet.com/news/smule-is-the-biggest-music-app-you-never-heard-of/

START (National Consortium for the Study of Terrorism and Responses to Terrorism). (2018). *The Use of Social Media by United States Extremists.* University of Maryland. Récupéré sur https://www.start.umd.edu/pubs/START_PIRUS_UseOfSocialMediaByUSExtremists_ResearchBrief_July2018.pdf

Statista. (2019, décembre 31). *MAU of iQiyi's mobile app in China 2016-2019 Published by Lai Lin Thomala, Jun 3, 2020. In 2019, the video app of iQiyi reached an average of 476 million monthly active users. Founded in Beijing in 2010, Baidu's iQiyi is one of the largest online video platforms in the world with over 100 million subscribers.* Récupéré sur https://www.statista.com/statistics/1106091/china-online-video-platform-iqiyi-mobile-app-monthly-active-user-number/

Statista. (2019, août 9). *Number of global monthly active Kakaotalk users from 1st quarter 2013 to 1st quarter 2019.* Récupéré sur

https://www.statista.com/statistics/278846/kakaotalk-monthly-active-users-mau/

Stokel-Walker, C. (2020, mars 20). *As humans go home, Facebook and YouTube face a coronavirus crisis*. Récupéré sur Wired: https://www.wired.co.uk/article/coronavirus-facts-moderators-facebook-youtube

Tardi, C. (2019, août 27). Monthly Active User (MAU). *Investopedia*. Récupéré sur https://www.investopedia.com/terms/m/monthly-active-user-mau.asp

Tech Against Terrorism. (2019, avril). *Analysis: ISIS use of smaller platforms and the DWeb to share terrorist content – April 2019*. Récupéré sur https://www.techagainstterrorism.org/2019/04/29/analysis-isis-use-of-smaller-platforms-and-the-dweb-to-share-terrorist-content-april-2019/

Tech Against Terrorism. (2020). *The Online Regulation Series - India*. Récupéré sur https://www.techagainstterrorism.org/2020/10/09/the-online-regulation-series-india/

Tech Against Terrorism. (2020). *The Online Regulation Series - Singapore*. Récupéré sur https://www.techagainstterrorism.org/2020/10/05/the-online-regulation-series-singapore/

Tech Against Terrorism. (2020). *The Online Regulation Series: The United States*. Récupéré sur Tech Against Terrorism: https://www.techagainstterrorism.org/2020/10/13/the-online-regulation-series-the-united-states/

Telegram. (s.d.). *ISIS Watch*. Récupéré sur Telegram: https://telegram.me/ISISwatch

Telegram. (s.d.). *Telegram Privacy Policy*. Récupéré sur Telegram: https://telegram.org/privacy?setln=fr

Tencent. (s.d.). *Agreement on Software License and Service of Tencent Weixin*. Récupéré sur https://weixin.qq.com/cgi-bin/readtemplate?lang=en&t=weixin_agreement&s=default&cc=CN

The Hindu Business Line. (2020, février 13). *Social media users to be tracked by government under new guidelines*. Récupéré sur https://www.thehindubusinessline.com/info-tech/social-media/social-media-users-to-be-tracked-by-government-under-new-guidelines-report/article30807839.ece

The International Centre for the Study of Radicalisation (ICSR). (2020). *ICSR info*. Récupéré sur The International Centre for the Study of Radicalisation (ICSR): https://icsr.info/

The Santa Clara Principles. (s.d.). *The Santa Clara Principles on Transparency and Accountability in Content Moderation*. Récupéré sur https://santaclaraprinciples.org

Thompson, E. (2021, janvier 29). *Canada not exempt from social media forces that created U.S. Capitol riot, heritage minister says*. Récupéré sur https://www.cbc.ca/news/politics/facebook-twitter-canada-regulation-1.5894301

Thune, J. (2020, juillet 29). *Thune: PACT Act Would Increase Internet Accountability and Consumer Transparency*. Récupéré sur https://www.thune.senate.gov/public/index.cfm/2020/7/thune-pact-act-would-increase-internet-accountability-and-consumer-transparency

TikTok. (2019-2020). *TikTok Transparency Report*. Récupéré sur https://www.tiktok.com/safety/resources/transparency-report?lang=en

Titcomb, J. (2017, novembre 1). *Why Google is reading your Docs*. Récupéré sur The Telegraph: https://www.telegraph.co.uk/technology/2017/11/01/google-reading-docs/

Tumblr. (2019). *Tumblr Government Transparency Report*. Récupéré sur Tumblr: https://static.tumblr.com/elwkrsl/u9Cqct0vc/government_transparency_report_2019.pdf

Twitch. (2020). *Rapport de transparence 2020*. Récupéré sur https://www.twitch.tv/p/fr-fr/legal/transparency-report/

Twitch. (s.d.). *Utilisation d'AutoMod*. Récupéré sur Twitch TV: https://help.twitch.tv/s/article/how-to-use-automod?language=fr

Twitter. (2012-2020). *Twittter Rules enforcement*. Récupéré sur Twitter Transparency Report: https://transparency.twitter.com/en/twitter-rules-enforcement.html

Twitter. (s.d.). *Notre approche en matière d'élaboration de politiques et notre philosophie relative à leur application*. Récupéré sur Twitter: https://help.twitter.com/fr/rules-and-policies/enforcement-philosophy

Twitter. (s.d.). *Notre gamme d'options pour l'application de nos politiques*. Récupéré sur help.twitter.com: https://help.twitter.com/fr/rules-and-policies/enforcement-options

US Treasury. (2020, janvier 23). *OFFICE OF FOREIGN ASSETS CONTROL - Specially Designated Nationals and Blocked Persons List.* Récupéré sur Treasury: https://www.treasury.gov/ofac/downloads/sdnlist.pdf

Verizon Media. (2019). *Transparency Report*. Récupéré sur Verizon Media: https://www.verizonmedia.com/transparency/index.html?guce_referrer=aHR0cHM6Ly90cmFuc3BhcmVuY3kub2F0aC5jb20vaW5kZXguaHRtbD9ndWNlX3JlZmVycmVyPWFIUjBjSE02THk5M2QzY3VkSFZ0WW14eUxtTnZiUzgmZ3VjZV9yZWZlcnJlcl9zaWc9QVFBQUFKazduZ3VNWS04dHhtNG9hWFM3TUlkNkxlUWxkMEZ5

Vimeo. (s.d.). *Comment Vimeo gère-t-il les contenus à caractère violent ? - Centre d'aide*. Récupéré sur Vimeo Zendesk: https://vimeo.zendesk.com/hc/fr/articles/224822427-Comment-Vimeo-g%C3%A8re-t-il-les-contenus-%C3%A0-caract%C3%A8re-violent-

VK. (2020). *Centre de sécurité*. Récupéré sur VK: https://m.vk.com/safety?section=social&lang=fr

VK. (2020). *Platform Standards*. Récupéré sur https://m.vk.com/safety?lang=en&section=standarts

Wang, Z. (2017). Systematic Government Access to Private-Sector Data in China. Dans F. H. Dempsey (Éd.), *Bulk Collection - Systematic Government Access to Private-Sector Data.* Oxford University Press.

Weimann, G. (2014). *New Terrorism and New Media.* Commons Lab of the Woodrow Wilson International Center for Scholars. Récupéré sur https://www.wilsoncenter.org/publication/new-terrorism-and-new-media

Wickey, W. (2018, août 23). *Should You Use Medium As Your Business Blog Platform? [2019 Update].* Récupéré sur Medium: https://medium.com/crowdbotics/medium-business-blog-platform-b8b8faa2d430

Wikimedia Foundation. (2019, juin 7). *Conditions d'utilisation - Wikimedia Foundation - Governance Wiki*. Récupéré sur Wikimedia Foundation: https://foundation.wikimedia.org/wiki/Terms_of_Use/fr

Wikimedia Foundation. (s.d.). *Transparency report*. Récupéré sur Wikimedia Foundation: https://transparency.wikimedia.org/

Wikipédia. (2019, décembre 14). *Core Content Policies - Wikipédia*. Récupéré sur Wikipédia: https://en.wikipedia.org/wiki/Wikipedia:Core_content_policies

Wikipédia. (2019, octobre 22). *Wikipédia: Vérificateur d'adresses IP*. Récupéré sur Wikipédia: https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:V%C3%A9rificateur_d%27adresses_IP

Wikipédia. (2020, janvier 1). *Administration - Wikipédia*. Récupéré sur Wikipédia: https://en.wikipedia.org/wiki/Wikipedia:Administration#Human_and_legal_administrati

on

Wikipedia. (2020, janvier 23). *Deletion process - Wikipédia*. Récupéré sur Wikipédia: https://en.wikipedia.org/wiki/Wikipedia:Deletion_process

Wikipédia. (2020, janvier 27). *What Wikipedia is not - Wikipéia*. Récupéré sur Wikipédia.

Wikipédia. (2020, janvier 26). *Wikipédia :Critères de suppression immédiate*. Récupéré sur Wikipédia: https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Crit%C3%A8res_de_suppression_imm%C3%A9diate

Wikipédia. (2020, janvier 28). *Wikipédia: Masqueur de modifications*. Récupéré sur Wikipédia: https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Masqueur_de_modifications

WordPress. (s.d.). *Terrorist Activity - Support - Word Press.com*. Récupéré sur WordPress: https://en.support.wordpress.com/terrorist-activity/

WordPress.com. (s.d.). *Contenu et sites suspendus*. Récupéré sur WordPress.com: https://wordpress.com/fr/support/blogs-suspendus/

Yahoo! Finance. (2019, novembre 13). *YY earnings surpass estimates in Q3, revenues increase*. Récupéré sur Yahoo! Finance: https://finance.yahoo.com/news/yy-earnings-surpass-estimates-q3-144502223.html

Yoo, E. (2018, avril 13). *Huoshan latest video platform to clean up vulgar content*. Récupéré sur technode: https://technode.com/2018/04/13/huoshan-clean-up/

Youku Tudou Inc. (NYSE: YOKU). (s.d.). *Youku Tudou Inc. (NYSE: YOKU), About us - 优酷视频*. Récupéré sur c.you.ku.com: https://c.youku.com/abouteg/youtu

YouTube. (2020, mars 16). *Protecting our extended workforce and the community*. Récupéré sur YouTube Official Blog: https://blog.youtube/news-and-events/protecting-our-extended-workforce-and

YY Inc. - IR Site. (2019, mai 28). *YY Reports First Quarter 2019 Unaudited Financial Results.* Récupéré sur http://ir.yy.com/news-releases/news-release-details/yy-reports-first-quarter-2019-unaudited-financial-results

Zetter, K. (2015, novembre 19). *Security Manual Reveals the OPSEC Advice ISIS Gives Recruits*. Récupéré sur Wired: https://www.wired.com/2015/11/isis-opsec-encryption-manuals-reveal-terrorist-group-security-protocols/

Zhong, R. (2018, novembre 8). *At China's Internet Conference, a Darker Side of Tech Emerges*. Récupéré sur The New York Times: https://www.nytimes.com/2018/11/08/technology/china-world-internet-conference.html

Notes

[1] Voir la section 1 des profils des services à l'Annexe B.

[2] Voir les sections 5 et 6 des profils des services à l'Annexe B.

[3] « Le nombre d'utilisateurs actifs mensuels renseigne sur la santé globale d'une entreprise en ligne et est utilisé pour le calcul d'autres indicateurs relatifs aux sites web. Il est également utile pour mesurer l'efficacité des campagnes de marketing d'une entreprise, ainsi que l'expérience des clients et prospects. Les investisseurs spécialisés dans le secteur des médias sociaux surveillent de près ce chiffre lorsque les entreprises le publient, car cet [indicateur de performance clé] peut influer sur le cours de bourse d'une entreprise de médias sociaux. » (Tardi, 2019)

[4] Voir les profils des services à l'Annexe [B] du premier rapport d'analyse comparative.

[5] Les informations issues des médias et autres sources librement accessibles ont en revanche été utilisées pour la section 10 de chaque profil (voir Annexe B), principalement parce que les documents constitutifs des services mentionnent rarement des incidents concrets liés à l'exploitation de leurs technologies à des fins terroristes et extrémistes violentes. En tout état de cause, ces sources, lorsqu'elles sont utilisées, sont dûment référencées dans les notes de bas de page.

[6] Facebook, YouTube, TikTok, Twitter et Google Drive.

[7] Voir la section 1 des profils de Facebook, YouTube, TikTok, Twitch, Twitter et Google Drive. On peut considérer que Microsoft (LinkedIn, Skype et OneDrive) en fait partie, bien que l'entreprise ne donne pas de définition de l'extrémisme violent et ne fournisse pas d'exemple. Inversement, Discord fournit des explications et des descriptions utiles de l'extrémisme violent et de l'incitation à la haine, mais ne définit pas le terrorisme. Pinterest propose également de bonnes descriptions des activités et des contenus haineux, mais ne définit pas les notions de terroriste et d'organisation terroriste.

[8] Instagram, Youku Tudou, iQIYI, LinkedIn, Baidu Tieba, Skype, Quora, Snapchat, Pinterest, Ask.fm, Xigua, Tumblr, Flickr, Huoshan, Haokan, Meetup, Dropbox, Microsoft OneDrive et Wordpress.com.

[9] Voir la section 1 des profils des services Instagram, Youku Tudou, iQIYI, Kuaishou, LinkedIn, Baidu Tieba, Skype, Quora, Snapchat, Pinterest, Ask.fm, Xigua, Discord, Tumblr, Flickr, Huoshan, Haokan, Meetup, Dropbox, Microsoft OneDrive et Wordpress.com.

[10] WeChat, Instagram, QQ, Youku Tudou, iQIYI, Douban, LinkedIn, Baidu Tieba, Vimeo, Twitch, Medium, Odnoklassniki, KaKaoTalk, Meetup et MySpace.

[11] Voir la section 1 des profils des services WeChat, Instagram, QQ, Youku Tudou, iQIYI, Kuaishou, Douban, LinkedIn, Baidu Tieba, Vimeo, Medium, Odnoklassniki et Meetup.

[12] WhatsApp, iMessage/FaceTime, QZone, Weibo, Reddit, Viber, IMO, Telegram, LINE, VK, YY Live, Discord, Smule, DeviantArt, 4chan et Wikipédia.

[13] Voir la section 1 des profils des services WhatsApp, iMessage/FaceTime, QZone, Weibo, Reddit, Viber,

IMO, Telegram, LINE, VK, YY Live, Smule, DeviantArt, 4chan et Wikipédia.

[14] Voir la section 1 des profils de Facebook et d'Instagram.

[15] Voir la section 7 du profil de YouTube ainsi que la section 1 des profils des services Skype, Quora, Microsoft OneDrive et Wordpress.com.

[16] Voir la section 1 du profil de VK.

[17] Voir la section 1 des profils des services WhatsApp, iMessage/Facetime, WeChat, QQ, Youku Tudou, Weibo, QZone, iQIYI, Reddit, Kuaishou, Telegram, Snapchat, Pinterest, Twitter, Douban, Baidu Tieba, Xigua, Viber, Discord, Vimeo, IMO, LINE, Huoshan, Ask.fm, YY Live, Twitch, Tumblr, Flickr, Medium, Odnoklassniki, Haokan Video, Smule, KakaoTalk, DeviantArt, Meetup, 4chan, Google Drive, Dropbox et Wikipédia.

[18] Le chiffrement préserve la confidentialité de la communication entre l'expéditeur et le destinataire, de sorte qu'aucun tiers ne peut accéder à cette communication, pas même l'entreprise fournissant le service. Le chiffrement protège également les informations conservées sur les ordinateurs, les téléphones mobiles et les autres appareils numériques, assurant la protection des informations sur l'appareil pour le cas où ce dernier serait perdu ou volé. Le chiffrement permet également aux personnes de s'exprimer librement, d'échanger des informations personnelles ou sensibles et de protéger leurs données. A fortiori, il permet également aux criminels de faire un usage abusif des systèmes de protection de la confidentialité et de la vie privée ainsi que des algorithmes de chiffrement de sécurité pour comploter et coordonner des attaques terroristes, mener des actions relevant du crime organisé et préserver leur anonymat. En conséquence, le chiffrement constitue un difficile compromis entre, d'une part, le respect de la vie privée et la sécurité et, d'autre part, l'application de la loi et l'établissement de rapports de transparence.

[19] Voir la section 2 – Différences entre les rapports de transparence publiés à l'heure actuelle sur les contenus terroristes et extrémistes violents du premier rapport d'analyse comparative.

[20] Facebook, YouTube, WhatsApp, Facebook Messenger, iMessage/FaceTime, Instagram, TikTok, Weibo, Reddit, Twitter, LinkedIn, Baidu Tieba, Skype, Quora, Snapchat, Viber, Pinterest, Vimeo, Telegram, LINE, Ask.fm, Xigua, Tumblr, Flickr, Houshan, VK, Medium, Odnoklassniki, Discord, Smule, KaKaoTalk, DeviantArt, Meetup, 4chan, MySpace, Google Drive, Dropbox, OneDrive, WordPress.com et Wikipédia.

[21] Voir la section 5 des profils de Facebook, YouTube, WhatsApp, Facebook Messener, iMessage/FaceTime, Instagram, TikTok, Weibo, Reddit, Kuaishou,Twitter, LinkedIn, Baidu Tieba, Skype, Quora, Snapchat, Viber, Pinterest, Vimeo, Telegram, LINE, Ask.fm, Xigua, Tumblr, Flickr, Houshan, VK, Medium, Odnoklassniki, Discord, Smule, KaKaoTalk, DeviantArt, Meetup, 4chan, Google Drive, Dropbox, OneDrive, WordPress.com et Wikipédia.

[22] Reddit, Viber, Twitch, Flickr, VK, Odnoklassniki, KaKaoTalk, DeviantArt, 4chan et Wikipédia.

[23] Voir les sections 4 et 5 des profils de Reddit, Viber, Twitch, Flickr, VK, Odnoklassniki, KaKaoTalk, DeviantArt, 4chan et Wikipédia.

[24] L'expression « au moins » est indiquée ici, car il n'est pas possible de déterminer à partir des informations librement accessibles le type d'activité et de processus mis en place par les services pour veiller au respect de leurs conditions d'utilisation et des autres documents constitutifs.

[25] Facebook, YouTube, WhatsApp, Facebook Messenger, WeChat, Instagram (membre du Hash Sharing Consortium), TikTok, Reddit (membre du Hash Sharing Consortium), Twitter, LinkedIn (membre du Hash Sharing Consortium), Skype (membre indirect du GIFCT via Microsoft), Snapchat (membre du Hash Sharing Consortium), Pinterest (membre du GIFCT), LINE, Ask.fm (membre du Hash Sharing Consortium), Twitch (membre indirect du GIFCT via Amazon), VK, YY Live, Google Drive, Dropbox (membre du GIFCT) et OneDrive (membre du GIFCT).

[26] Ici encore, l'expression « au moins » est indiquée, car il n'est pas possible de déterminer à partir des informations librement accessibles le type d'activité et de processus mis en place par les services pour veiller au respect de leurs conditions d'utilisation et des autres documents constitutifs. Voir par exemple

la section 5 des profils de QQ, Youku Tudou, QZone TikTok, Weibo, iQIYI, Douban, Baidu Tieba, YY Live, Xigua, Huoshan et Haokan.

[27] Voir la section 5 des profils de Facebook, YouTube, WhatsApp, Facebook Messenger, WeChat, Instagram (membre du GIFCT), TikTok, Reddit (membre du Hash Sharing Consortium), Twitter, LinkedIn (membre du Hash Sharing Consortium), Skype (membre indirect du GIFCT via Microsoft), Snapchat (membre du Hash Sharing Consortium), Pinterest (membre du GIFCT), Viber, Discord (membre du GIFCT), LINE, Ask.fm (membre du Hash Sharing Consortium), Twitch (membre indirect du GIFCT via Amazon), VK, YY Live, Google Drive, Dropbox (membre du GIFCT) et OneDrive (membre du GIFCT).

[28] Facebook, YouTube, Facebook Messenger, Instagram, Reddit, Twitter, Quora, Pinterest, Vimeo, Ask.fm, Twitch, Tumblr, VK, Medium, Odnoklassniki, Smule, KaKaoTalk, DeviantArt, Meetup, Dropbox et Wordpress.com.

[29] Voir la section 4.1 des profils de Facebook, YouTube, Facebook Messenger, WhatsApp, Instagram, Reddit, Snapchat, Twitter, Quora, Pinterest, Vimeo, Ask.fm, Twitch, Tumblr, VK, Medium, Odnoklassniki, Smule, KaKaoTalk, DeviantArt, Meetup, Dropbox et Wordpress.com.

[30] Facebook, YouTube, WhatsApp, Facebook Messenger, Instagram, TikTok, Reddit, Twitter, Quora, Pinterest, Vimeo, LINE, Ask.fm, Twitch, Tumblr, VK, Medium, Discord, KaKaoTalk, DeviantArt, Meetup, 4chan et Wordpress.com.

[31] Voir la section 4.2 des profils de Facebook, YouTube, WhatsApp, Facebook Messenger, Instagram, TikTok, Reddit, Kuaishou, Twitter, Snapchat, Quora, Viber, Pinterest, Vimeo, LINE, Ask.fm, Twitch, Tumblr, VK, Medium, Discord, KaKaoTalk, DeviantArt, Meetup, 4chan, Dropbox et Wordpress.com.

[32] WhatsApp, iMessage/FaceTime, WeChat, Instagram, QQ, TikTok, Weibo, iQIYI, Douban, LinkedIn, Quora, Snapchat, Pinterest, IMO, Ask.fm, VK, Haokan, Odnoklassniki, Smule, Meetup, MySpace et OneDrive.

[33] Voir les sections 4 et 5 des profils de iMessage/FaceTime, WeChat, QQ, Weibo, iQIYI, Kuaishou, Douban, LinkedIn, Quora, Pinterest, IMO, Ask.fm, VK, Haokan, Odnoklassniki, Smule et Meetup. Le recours à des formulations de type « pourrait examiner... » ou « se réserve le droit d'examiner... », en particulier, est monnaie courante.

[34] Voir les sections 4 et 5 des profils de WeChat, QQ, Youku Tudou, QZone, Weibo, iQIYI, Kuaishou, Douban, Baidu Tieba, YY Live, Xigua, Huoshan et Haokan Video.

[35] Voir par exemple https://extremism.gwu.edu/sites/g/files/zaxdzs2191/f/EncryptedExtremism.pdf et https://www.counterextremism.com/sites/default/files/Extremists%20and%20Online%20Propaganda_040918.pdf

[36] http://www.terrorismanalysts.com/pt/index.php/pot/article/view/607/1200

[37] Voir la campagne de consultation et la documentation afférente à l'adresse https://www.communications.gov.au/have-your-say/consultation-bill-new-online-safety-act.

[38] Pour plus d'informations sur les contenus violents odieux et les programmes de blocage des fournisseurs de services internet en Australie, veuillez consulter les références suivantes :

- Blog eSafety sur la panoplie d'outils et de pouvoirs adoptés en réponse aux événements de Christchurch : https://www.esafety.gov.au/about-us/blog/christchurch-shifted-online-world-its-axis
- Fiche eSafety sur les contenus violents odieux : https://www.esafety.gov.au/sites/default/files/2020-03/eSafety-AVM-factsheet.pdf
- Fiche eSafety sur le blocage des fournisseurs de services internet : https://www.esafety.gov.au/sites/default/files/2020-03/eSafety-AVM-factsheet.pdf
- Communiqué de presse eSafety sur le protocole de blocage des principaux fournisseurs de services internet : https://www.esafety.gov.au/about-us/newsroom/blocking-viral-spread-terrorist-content-online

[39] Voir https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN

[40] Le texte de la nouvelle législation a été rendu disponible en anglais dans le cadre de la procédure européenne de notification : https://ec.europa.eu/growth/tools-databases/tris/en/index.cfm/search/?trisaction=search.detail&year=2020&num=65&mLang=EN

[41] De plus amples informations de même qu'un résumé de la législation NetzDG et des réponses aux questions les plus fréquemment posées sont disponibles en anglais à l'adresse https://www.bmjv.de/DE/Themen/FokusThemen/NetzDG/NetzDG_EN_node.html.

[42] Voir le texte sur https://www.gov.ie/en/publication/d8e4c-online-safety-and-media-regulation-bill/.

[43] Son texte est disponible à l'adresse http://www.legislation.govt.nz/bill/government/2020/0268/latest/LMS294551.html.

[44] Le texte visé est disponible à l'adresse https://www.legislation.govt.nz/act/public/2015/0063/latest/DLM5711810.html.

[45] Ce profil concerne la plateforme Facebook et non la société dans son ensemble. Il ne porte donc pas sur Messenger, Instagram ou WhatsApp.

[46] Le programme YouTube Trusted Flagger a été élaboré par YouTube pour mettre à la disposition de personnes, d'organismes publics et d'organisations non gouvernementales des outils robustes particulièrement efficaces pour signaler à YouTube les contenus qui enfreignent son règlement de la communauté. https://support.google.com/youtube/answer/7554338?&ref_topic=2803138

[47] Voir la section 3 du rapport.

[48] Il est à noter que ces conditions s'appliquent exclusivement aux utilisateurs de QQ, où qu'ils soient dans le monde, sauf s'ils appartiennent à l'une des catégories suivantes : (a) utilisateur de QQ en République populaire de Chine ; (b) ressortissant de la République populaire de Chine utilisant QQ en quelque point du globe ; (c) société de constitution chinoise utilisant QQ en un lieu quelconque dans le monde. Les utilisateurs appartenant à ces catégories sont soumis aux conditions d'utilisation applicables aux utilisateurs de la République populaire de Chine, disponibles à l'adresse https://www.qq.com/contract.shtml.

[49] Qzone n'est accessible hors de Chine qu'à travers QQ International.

[50] Ces conditions d'utilisation s'appliquent aux utilisateurs hors de Chine. Les utilisateurs du service QZone en Chine sont soumis aux conditions d'utilisation applicables aux utilisateurs de la République populaire de Chine, disponibles à l'adresse https://www.qq.com/contract.shtml.

[51] Quoique Tumblr ait indiqué qu'il participait au Hash Sharing Consortium, aucune mention n'en est faite sur le site internet du GIFCT (septembre 2020).