unesco

LES POLITIQUES DES GRANDES PLATEFORMES SUR LE DISCOURS DE HAINE PENDANT LA COVID-19

Ana Laura Pérez

Publié en 2021 par l'Organisation des Nations Unies pour l'éducation, la science et la culture, 7, place de Fontenoy, 75352 Paris 07 SP, France et le Bureau régional de l'UNESCO pour la science en Amérique latine et dans les Caraïbes, UNESCO Montevideo, Luis Piera 1992, 2ème étage, 11200 Montevideo, Uruguay.

MTD/CI/2021/PI/01/REV1

Coordination éditoriale : Sandra Sharman
Conception graphique : Trigeon.

Cette publication a été s
outenue par OBSERVACOM.

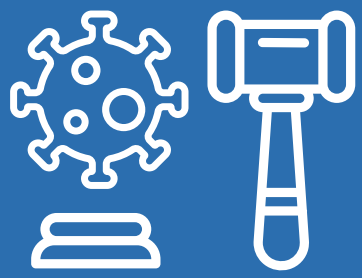
CONTENU

RÉSUMÉ ANALYTHIQUE

Ce document fait état d'une augmentation des messages considérés comme des discours de haine sur Facebook, Twitter et YouTube, depuis l'avènement de la pandémie de COVID-19. Bien qu'inégale, cette augmentation peut être établie à partir des rapports de transparence des différentes plateformes et de la croissance enregistrée dans la modération de ces contenus à partir de mars 2020.

Étant donné qu'au cours de la même période, et en conséquence des mesures d'isolement prises dans la plupart des pays du monde, les plateformes ont décidé d'accroître l'utilisation d'outils d'intelligence artificielle dans leurs processus de modération, il n'est pas possible d'être certain que cette croissance est due à une augmentation de la création et de la publication de messages ou à un changement dans les systèmes de détection qui a affecté les résultats d'une année à l'autre.

INTRODUCTION



L'arrivée de la pandémie de COVID-19 en 2020 a eu des répercussions bien au-delà des services de santé et des populations du monde entier. Dans certains cas, nous ne les avons pas encore pu analyser complètement, et il faudra probablement des années pour les déterminer avec certitude.

Certaines études font état d'une augmentation des contenus classables comme discours de haine au sein des plateformes à l'égard de groupes spécifiques à la suite de la pandémie de COVID-19, ainsi que des preuves d'une augmentation du nombre de contenus retirés des réseaux sociaux sous cette étiquette.

Depuis 2020, les plateformes et réseaux sociaux ont apporté des modifications substantielles à leurs critères de modération, ont ajouté de nouvelles clauses à leurs normes communautaires et ont dû augmenter le poids de la modération automatique dans leurs processus réguliers, car un grand nombre de leurs employés sont rentrés chez eux. En outre, les préoccupations concernant l'impact des discours de haine sur ces plateformes et leurs conséquences possibles sur l'apparition de violences ont définitivement fait leur entrée dans le débat public.

Cette année, Twitter, Facebook et YouTube - à des degrés divers chacune - sont passées d'essayer de rester impartiales quant au débat public sur leurs plateformes à, par exemple, bloquer les comptes d'un président en exercice pendant les derniers jours de son administration.

En examinant les rapports de transparence des plateformes sociales, il est clair qu'au cours de l'année 2020, les contenus catégorisés comme discours de haine ont augmenté de manière significative sur les réseaux sociaux, tout comme la suppression de posts pour ces raisons. L'absence d'informations suffisamment désagrégées sur ce que chacune des plateformes analysées entend par discours de haine, ainsi que sur les processus décisionnels et les taux d'erreur dans la prise de décisions, rend difficile la détermination des raisons de cette croissance.

Pour leur part, les plateformes ont déclaré publiquement qu'elles avaient rencontré des problèmes dans les processus de modération comme conséquence de l'envoi chez eux de milliers de travailleurs affectés dans ces domaines et de la décision résultante d'augmenter le poids de la modération automatisée et des systèmes d'intelligence artificielle. Elles ont également reconnu la possibilité que cela ait entraîné une augmentation des taux d'erreur en raison des problèmes rencontrés par les logiciels d'apprentissage automatique pour comprendre le contexte dans lequel une grande partie du contenu est créée et les différences entre ces capacités et celles des modérateurs humains.

Facebook en particulier, tant cette plateforme que sa sœur Instagram, ont connu une augmentation exponentielle des contenus qualifiés de discours de haine durant la période où la COVID-19 s'est installée dans le monde. Il est notable que l'augmentation est enregistrée de manière significative à partir du deuxième trimestre de 2020, lorsque dans la plupart des pays du monde les gouvernements ont commencé à mettre en œuvre des mesures de distanciation physique soutenue et de quarantaine ou de confinement.

Cependant, aucune donnée ne permet d'établir si les raisons de cette augmentation peuvent également inclure un changement des critères d'examen des contenus et une évolution vers un modèle plus agressif de modération des contenus ou une augmentation réelle des discours de haine sur les réseaux sociaux à partir de 2020.

Cet article vise à analyser l'augmentation des discours de haine en ligne depuis l'arrivée de la pandémie de COVID-19 dans le monde et les actions mises en œuvre par Facebook, Twitter et YouTube, afin d'en définir la portée, les effets, les motifs et les conséquences possibles.

DISCOURS DE HAINE SUR INTERNET

Le discours de haine est un concept complexe, qui comporte définitions variables, sur lequel il n'y a pas d'accord entre les plateformes, les gouvernements et leurs réglementations et lois. Dans **la Stratégie et plan d'action des Nations unies pour la lutte contre les discours de haine, signés par le secrétaire général de l'ONU, António Guterres,** le discours de haine est défini comme « tout type de communication, qu'il s'agisse d'expression orale ou écrite ou de comportement, constituant une atteinte ou utilisant un langage péjoratif ou discriminatoire à l'égard d'une personne ou d'un groupe en raison de leur identité, en d'autres termes, de l'appartenance religieuse, de l'origine ethnique, de la nationalité, de la race, de la couleur de peau, de l'ascendance, du genre ou d'autres facteurs constitutifs de l'identité». On ajoute que « souvent, ces discours sont à la fois le résultat et la cause de l'intolérance et de la haine et peuvent être, dans certains cas, dénigrants et source de divisions ».

D'un point de vue juridique, le droit international n'interdit pas le discours de haine en tant que tel mais « l'incitation à la discrimination, à l'hostilité et à la violence», la première étant définie par les Nations unies comme « une forme discursive particulièrement dangereuse car elle vise explicitement et délibérément à provoquer des actes de discrimination, d'hostilité ou de violence, et peut également conduire à la commission d'attentats terroristes ou d'atrocités criminelles ». Pour cette raison, explique le document, le droit international n'oblige pas les États à interdire les discours de haine qui n'atteignent pas le seuil de l'incitation. Mais l'ONU prévient néanmoins que « même s'ils ne sont pas interdits, ces propos peuvent être préjudiciables ».

« Dans le monde entier, nous assistons au déferlement de la xénophobie, du racisme et de l'intolérance ; cette tendance alarmante va notamment de pair avec la montée de l'antisémitisme, de la haine à l'égard des musulmans et de la persécution des chrétiens. Les médias sociaux et d'autres moyens de communication servent de tribunes au fanatisme. Les mouvements néonazis et de la suprématie blanche sont de plus en plus nombreux. Les débats publics utilisent une rhétorique incendiaire à des fins politiques pour stigmatiser et déshumaniser les minorités, les migrants, les réfugiés et toute personne qu'on dit « autre ». Il ne s'agit ni d'un phénomène isolé ni des hauts cris de quelques-uns qui vivent en marge de la société. La haine prend ses quartiers sur la place publique, au sein des démocraties libérales comme des régimes autoritaires. Et chaque fois qu'une norme n'est plus respectée, ce sont tous les piliers de notre humanité commune qui vacillent. Les discours de haine constituent une menace pour les valeurs démocratiques, la stabilité sociale et la paix », assure l'ONU dans le document.

Également, dans la **Recommandation de Politique Générale N° 15 sur la lutte contre le discours de haine de la Commission européenne contre le racisme et**

l'intolérance (ECRI) du Conseil de l'Europe, le discours de haine est défini comme « l'usage d'une ou de plusieurs formes particulières d'expression – à savoir, l'appel à, la promotion de ou l'incitation au dénigrement, à la haine ou à la diffamation à l'encontre d'une personne ou d'un groupe de personnes, ainsi que le harcèlement, les injures, les stéréotypes négatifs, la stigmatisation ou les menaces à l'encontre de cette ou ces personne(s) et toute justification de ces diverses formes d'expression – fondée(s) sur une liste non exhaustive de caractéristiques ou de situations personnelles englobant la «race», la couleur de peau, la langue, la religion ou les convictions, la nationalité ou l'origine nationale ou ethnique ainsi que l'ascendance, l'âge, un handicap, le sexe, le genre, l'identité de genre et l'orientation sexuelle ».

Selon la définition, les discours de haine « ne visent pas seulement à inciter à la commission d'actes de violence, d'intimidation, d'hostilité ou de discrimination mais dont on peut raisonnablement attendre qu'ils aient cet effet » et « aux motifs autres que la « race», la couleur de peau, la langue, la religion ou les convictions, la nationalité ou l'origine nationale ou ethnique ainsi que l'ascendance». Il est également ajouté que la portée du terme «expression » s'entend comme «englobant les discours et les publications de toute forme, notamment par le biais des médias électroniques, ainsi que leur diffusion et leur conservation », ainsi que le fait que le discours de haine « peut prendre la forme de propos écrits ou exprimés de vive voix ou d'autres formes telles que des images, des signes, des symboles, des peintures, de la musique, des pièces de théâtre ou des vidéos » et « recouvre également l'adoption d'un comportement particulier (des gestes par exemple) pour communiquer une idée, un message ou une opinion ». La définition comprend « la négation, la banalisation, la justification ou l'apologie publiques de crimes de génocide, de crimes contre l'humanité ou de crimes de guerre confirmés par la justice et l'éloge des personnes ayant commis ces crimes ».

Plusieurs pays dans le monde disposent d'une législation interdisant les discours de haine qui se concentrent généralement sur l'incitation à la haine envers des personnes en raison de leurs caractéristiques identitaires.

En Amérique latine, l'approche a eu tendance à être très axée sur la législation et, comme l'indique Marianne Díaz Hernández dans **son ouvrage Discours de haine en Amérique latine : tendances de réglementation, rôle des intermédiaires et risques pour la liberté d'expression** , dans la plupart des cas ils misent sur les sanctions pénales directes, les sanctions pénales accessoires (en tant que facteur aggravant d'une infraction principale) et l'interdiction qui, sans créer de sanctions pénales, établit des mesures de réparation. Díaz Hernández ajoute qu'en Amérique latine, plusieurs pays (le Costa Rica, El Salvador, le Pérou, l'Argentine, la Bolivie et l'Uruguay, pour n'en citer que quelques-uns) « ont criminalisé l'incitation à la haine en tant que crime dans leur législation pénale générale ».

Par ailleurs, parmi les pays qui ont choisi le modèle punitif, tous ne caractérisent pas l'incitation à la haine selon les mêmes paramètres, certains exigeant la présence réelle ou potentielle d'un préjudice pour constituer le crime. À cet égard, la

Commission interaméricaine des droits de l'homme a souligné que « par principe, plutôt que de les restreindre, les États devraient encourager des mécanismes de prévention et d'éducation et promouvoir des débats plus larges et plus profonds, en tant que mesure visant à exposer et à combattre les stéréotypes négatifs ».

Cependant, il existe un certain consensus sur le fait que le discours de haine peut jouer un rôle dans la création de conditions propices à la violence envers des groupes spécifiques de la société. L'académicien Alexander Tsesis **argues that the v affirme que la principale motivation des discours de haine intimidants est de perpétuer et d'accroître les inégalités existantes.**

"« Bien que la diffusion des discours de haine intimidants n'entraîne pas toujours une violence discriminatoire, elle établit un raisonnement pour le ciblage de groupes particulièrement défavorisés », affirme-t-il.

Les actes de violence perpétrés contre les Rohingyas au Myanmar, par exemple, montrent le rôle qu'ont joué les publications sur Facebook contenant des discours de haine. En 2018, **une enquête de Reuters menée conjointement avec le Centre des droits de l'homme de la faculté de droit de l'UC Berkeley** a trouvé plus de 1 000 posts définissant les Rohingyas ou d'autres musulmans comme des chiens, des fumiers et des violeurs.

Ce contenu a été créé et diffusé au début d'une campagne de nettoyage ethnique et de crimes contre l'humanité menée par l'armée du Myanmar, qui a entraîné la fuite de 740 000 Rohingyas vers le Bangladesh.

2020 : UNE «AVALANCHE DE HAINE ET DE XÉNOPHOBIE »

En mai 2020, le secrétaire général de l'ONU, Antonio Guterres, a averti que la pandémie de COVID-19 avait généré une « avalanche de haine et de xénophobie » dans le monde: « on désigne des boucs émissaires; on entretient la peur ». Il a appelé à « agir maintenant pour renforcer l'immunité de nos sociétés face au virus de la haine ».

Il a dit que « des migrants et des réfugiés ont été accusés de propager le virus et se sont vus refuser l'accès aux soins médicaux», et il a ajouté que les personnes âgées ont été dépeintes comme des « caricatures méprisables » qui « suggèrent qu'elles sont aussi les plus sacrifiables ». Enfin, des journalistes, des professionnels de santé, des travailleurs humanitaires et des défenseurs des droits humains « sont pris pour cible simplement parce qu'ils font leur métier ».

À cet égard, la Haute-Commissaire des Nations Unies aux droits de l'homme, Michelle Bachelet, **a déclaré lors de la 13e session du Forum sur les questions relatives aux minorités en novembre 2020** que les médias sociaux ont représenté de nouvelles « opportunités pour l'exercice des libertés fondamentales telles que l'expression, l'association et la participation, les élargissant à des degrés sans précédent»; toutefois, « cette expansion a entraîné de nouvelles et importantes menaces pour l'espace civique et les droits des individus ».

« L'un d'entre eux est le discours de haine, qui est répandu en ligne sur diverses plateformes sociales. Les minorités ont été ciblées de manière disproportionnée pour incitation à la discrimination, à l'hostilité et à la violence. Cela peut entraîner des tensions, de l'agitation et des attaques contre des individus et des groupes. Cela peut également être utilisé pour servir des intérêts politiques et contribuer à créer un climat de peur au sein des communautés minoritaires », a déclaré Mme Bachelet.

La Haute-Commissaire a déclaré que « les droits dont jouissent les personnes hors ligne doivent également être protégés en ligne» et a ajouté que les entreprises de médias sociaux « ont la responsabilité de prévenir, d'atténuer et de réparer les violations des droits de l'homme qu'elles causent ou auxquelles elles contribuent ».

« Les entreprises de médias sociaux ont le choix de supprimer ou de laisser du matériel en ligne. Elles peuvent également marquer le contenu, ajouter des éléments compensatoires, avertir le diffuseur et suggérer une modération. L'élimination ne serait justifiée que dans les cas les plus graves. Toute solution proposée pour lutter contre les discours de haine sur les médias sociaux devrait s'efforcer de combler un énorme manque de transparence et de responsabilité démocratique dans la prise de

décisions des plateformes. Non seulement nous devons attendre d'elles qu'elles respectent les directives en matière de droits de l'homme, mais nous avons également besoin de mécanismes pour contrôler et évaluer leurs actions », a déclaré Mme Bachelet.

Dans le même ordre d'idées, le Rapporteur spécial de l'ONU sur la liberté de religion ou de conviction, Ahmed Shaheed, a dénoncé en avril 2020 la diffusion sur les plateformes sociales d'une théorie du « complot » prétendant « que les Juifs ou Israël sont responsables de la création et de la propagation du virus COVID-19 ».

« La lutte contre les discours de haine en ligne ne sera pas couronnée de succès si les médias généralistes ou les médias sociaux ne prennent pas au sérieux les rapports de cyber haine dirigés contre les Juifs et d'autres minorités (...). Les médias doivent supprimer tout message incitant à la haine ou à la violence, en plus d'identifier et de signaler les fausses nouvelles », a-t-il déclaré. Shaheed a ajouté que « en ces temps extraordinairement difficiles, il est plus nécessaire que jamais de veiller à ce que toutes les personnes puissent exercer, sans crainte et dans toute la mesure du possible, leur droit à la liberté de religion ou de conviction, tout en protégeant la santé publique ».

Pendant la pandémie, on a constaté une augmentation des discours haineux sur les médias sociaux. Tout d'abord, en février, la cible principale était la communauté chinoise, car c'est en Chine que le COVID-19 est né. Ensuite, la haine s'est portée sur l'utilisation de masques faciaux, allant même jusqu'à accuser la population LGBTIQ d'être l'origine supposée d'un virus considéré comme une punition divine. Selon **une étude menée par la société L1ght** fles discours de haine à l'égard de la Chine ou des ressortissants chinois ont augmenté de 900 % sur Twitter et le trafic vers les sites qui propagent des discours de haine ou vers des messages spécifiques à l'encontre de la communauté chinoise ou asiatique a augmenté de 200 %.

Dans de nombreux cas, ces expressions provenaient également de dirigeants politiques de différentes régions du monde, à la fois sur leurs plateformes sociales (avec des millions de followers) et en dehors de celles-ci. **L'utilisation du terme « virus chinois »** sur ses comptes de médias sociaux par le président américain de l'époque, Donald Trump, et l'utilisation de « virus de Wuhan » par le secrétaire d'État de l'époque, Mike Pompeo, ont peut-être encouragé **l'utilisation de discours haineux aux États-Unis.**

En février 2020, le gouverneur de la région italienne de Vénétie, Luca Zaia, l'un des premiers épicentres de la pandémie, a déclaré à des journalistes que le pays gérerait mieux le virus que la Chine en raison de **« l'hygiène de notre peuple (...) les citoyens italiens, la formation culturelle que nous avons, pour se doucher, se laver, se laver les mains très souvent (...), alors que nous avons tous vu les vidéos avec des Chinois mangeant des rats vivants ».**

En avril de la même année, **Abraham Weintraub** alors ministre brésilien de l'éducation, a suggéré dans un tweet que la pandémie faisait partie du « plan de domination mondiale » du gouvernement chinois.

Cette intensification de la rhétorique raciste sur les réseaux sociaux et dans les médias coïncide avec une augmentation des attaques contre ces mêmes groupes dans diverses régions du monde. **Au Royaume-Uni, des Asiatiques ont été battus** et sont devenus la **cible de railleries** et d'accusations d'avoir propagé le coronavirus. Deux femmes ont attaqué des étudiantes chinoises en Australie, les ont battues et ont donné des coups de pied à l'une d'entre elles en criant **« retournez en Chine »** et **« sales immigrantes ».** lEn Espagne, deux hommes ont battu **un jeune américain d'origine chinois** jusqu'à le plonger dans le coma pendant deux jours. Dans l'État américain du Texas, un homme armé d'un couteau **a attaqué une famille birmane,** l'accusant d'être un facteur de transmission du coronavirus.

En Afrique, des incidents de discrimination et des attaques contre des Asiatiques accusés d'être porteurs de coronavirus, ainsi que contre des étrangers en général, ont été signalés **au Kenya, en Éthiopie** et en **Afrique du Sud.**

Des cas ont également été signalés en Amérique latine. Au Brésil, les médias ont rapporté des **cas de harcèlement** et de rejet de personnes d'origine asiatique. Dans un de ces cas, une étudiante en droit a dénoncé avoir été victime de racisme et de xénophobie de la part d'une passagère du métro de Rio de Janeiro. « Cette femme a attendu que je me rende à la porte de la voiture pour crier 'regarde la chinoise qui s'en va, cochonne de chinoise', 'dégoûtante' et 'elle reste ici et nous rend tous malades' », a posté Marie Okabayashi sur Twitter avec une vidéo de l'agresseuse.

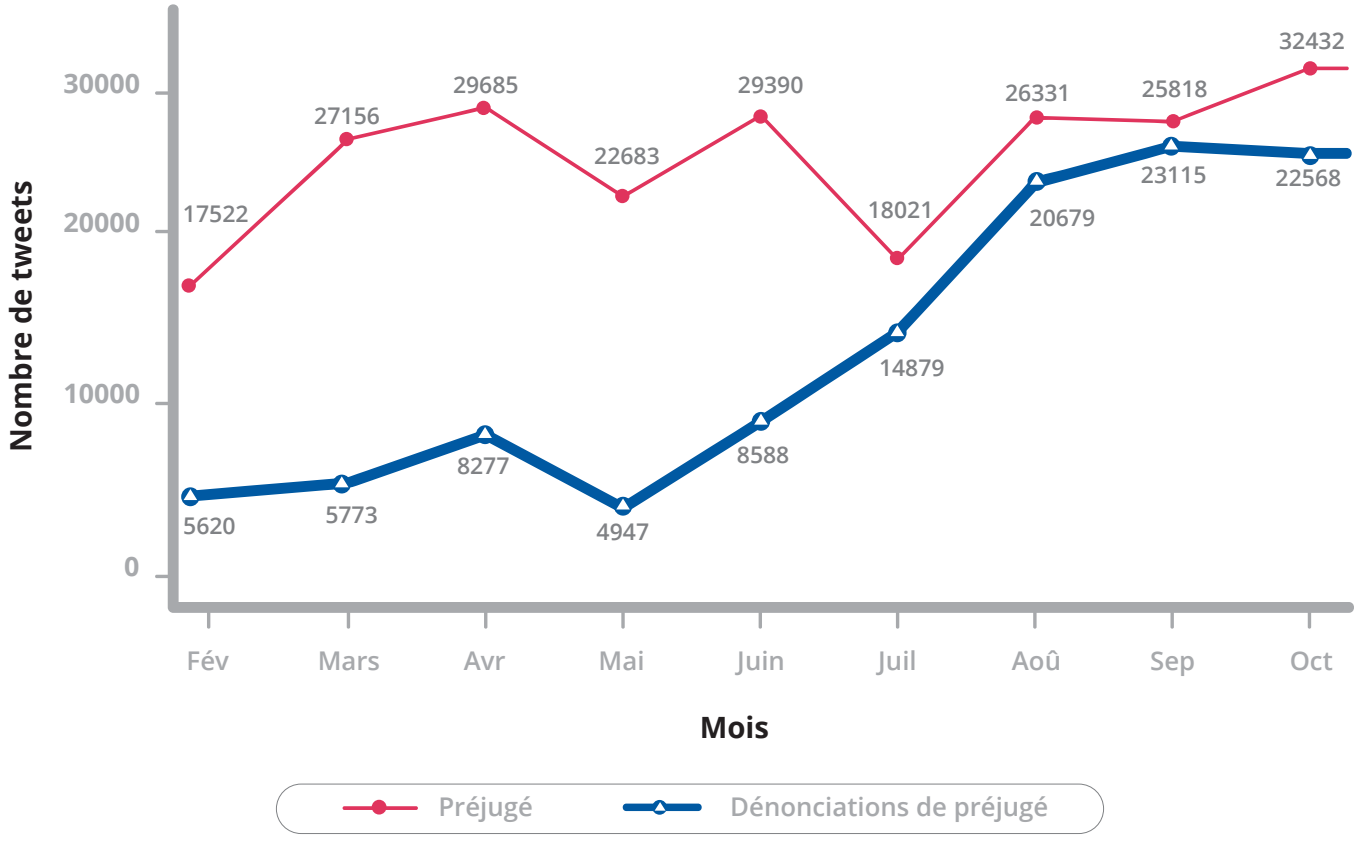
L'historienne mexicaine Yuriko Valdez, d'origine chinoise et auteur du documentaire El legado de mi raza. Chinos y mestizos en Mexicali (L'héritage de ma race. Chinois et métis à Mexicali), met en garde contre la prolifération d'attitudes xénophobes de la part de la communauté qui s'y trouve, ainsi que contre les nombreux commentaires racistes sur les réseaux sociaux dans des publications qui parlaient de célébrations comme le Nouvel An chinois du 25 janvier. Outre les commentaires habituels du type «les Chinois mangent des rats et des chiens», il y avait ceux du type « porcs chinois » ou « ils vont nous infecter parce que la Chine est la source d'infection du coronavirus ». Des réactions allant dans le même sens, de la part de « personnes fières qui prétendent être vraiment de Mexicali », sont apparues lors de la promotion de l'inauguration d'une exposition de l'association chinoise au zoo Bosque de la Ciudad : « Les Chinois ne méritent pas un hommage », « ils sont malades du coronavirus », entre autres messages que Valdez raconte.

« Les expressions de racisme et de xénophobie liées à la COVID-19 sur les plateformes numériques ont inclus le harcèlement, les discours de haine, la

prolifération de stéréotypes discriminatoires et les théories du complot. Il n'est pas surprenant que les leaders qui tentent d'attribuer la COVID-19 à certains groupes nationaux ou ethniques soient les mêmes leaders nationalistes populistes qui ont placé la rhétorique raciste et xénophobe au centre de leurs programmes politiques », a déclaré E. Tendayi Achiume, rapporteuse spéciale sur **les formes contemporaines de racisme, de discrimination raciale, de xénophobie et de l'intolérance qui y est associée.**

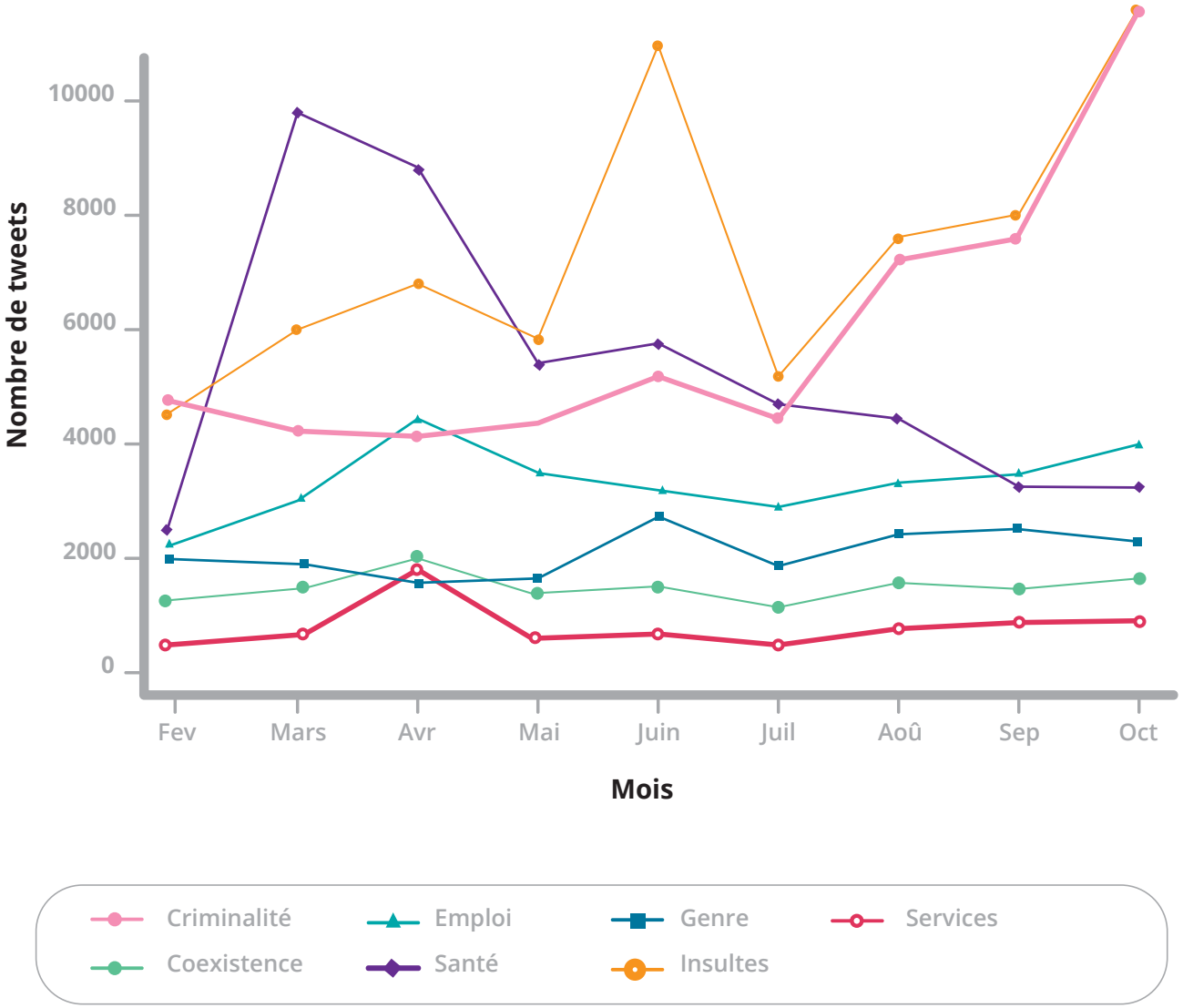
L'unité Migration de la Banque interaméricaine de développement (BID) a mené une étude entre février et décembre 2020 dans laquelle elle a suivi les conversations sur les **migrants sur Twitter**. La recherche a porté sur sept pays de la région qui sont considérés comme d'importants pays d'accueil de migrants : l'Argentine, le Chili, la Colombie, le Costa Rica, l'Équateur, le Panama et le Pérou. Dans ce suivi, des tweets ont été collectés avec des termes tels que asile, xénophobie, migrant, immigrant, réfugié, exilé, et une fois la collecte effectuée, un algorithme les a classés en huit catégories s'excluant mutuellement. Les sept premières catégories comprennent les tweets exprimant des préjugés envers les migrants dans les domaines de la criminalité, de l'emploi, du genre, des services sociaux, de la coexistence, de la santé et des insultes en général. La huitième catégorie englobe les tweets dénonçant ou répudiant ces préjugés, comme expliqué dans la recherche.

À partir de ces données, et en utilisant les tweets de février comme base de référence avant la pandémie, l'étude a révélé une augmentation des expressions de préjugés à l'égard des migrants de 70 % en deux mois, passant de 17 522 tweets mensuels en février à 29 685 en avril.



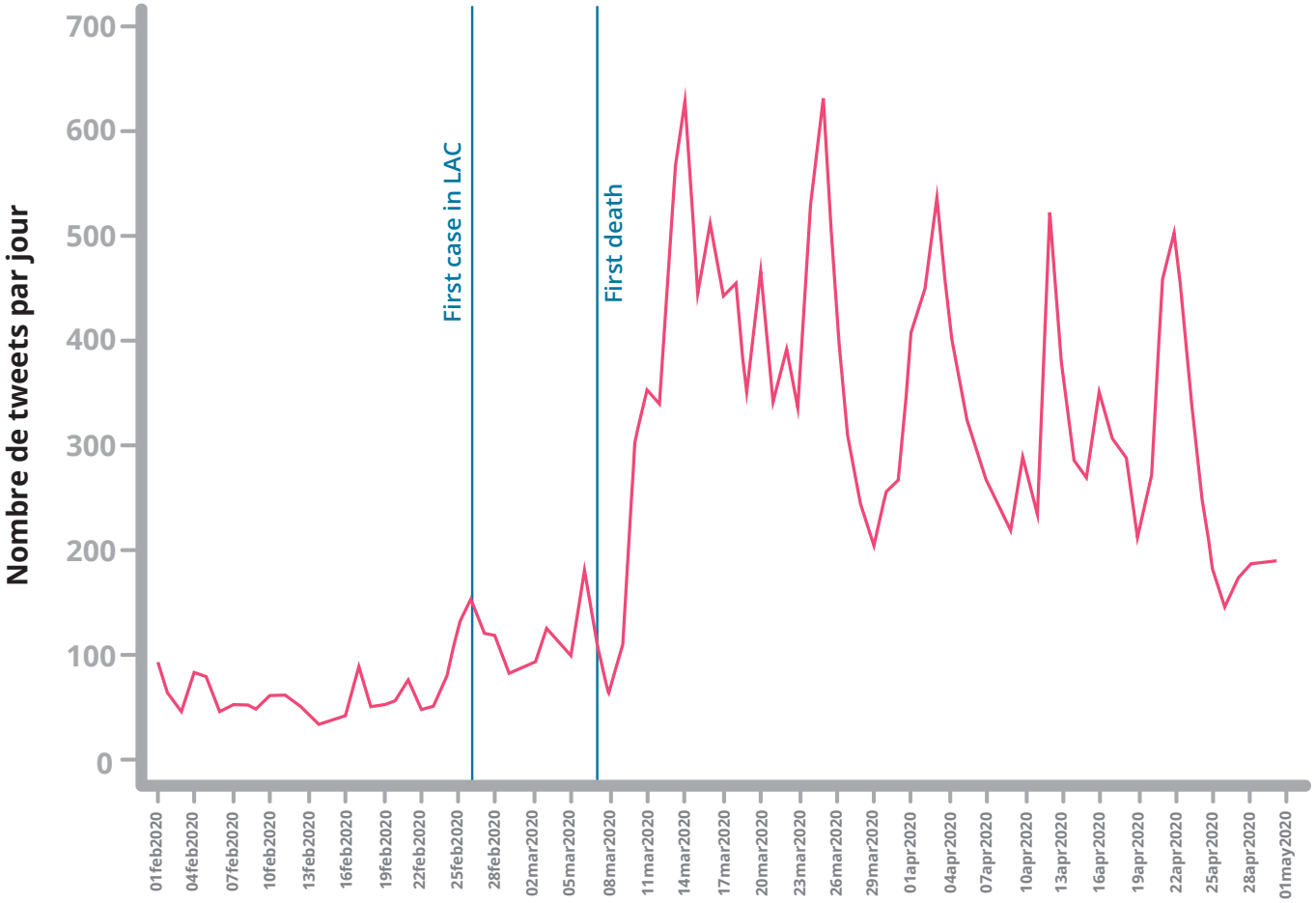
Source : Unité de migration de la BID, sur la base de données générée par Citibeats.

Selon l'étude, une grande partie de cette augmentation entre février et avril s'explique par des préjugés liés à la santé, « principalement expliqués par la crainte que les migrants ne transmettent la maladie ou ne fassent que les systèmes de santé s'effondrent ».



Source : Unité de migration de la BID, sur la base de données générée par Citibeats.

L'étude de la BID précise que « ces préjugés ont été déclenchés par le premier décès de COVID-19 annoncé dans la région », **survenu en mars 2020 en Argentine.**



Source : Unité de migration de la BID, sur la base de données générée par Citibeats.

Au cours des mois suivants, les auteurs de l'étude affirment avoir constaté des fluctuations dans les niveaux de xénophobie ou de préjugés, mais toujours plus élevés qu'en février avant la pandémie. En octobre, ils ont enregistré une hausse, expliquée dans ce cas par d'autres facteurs (comme la criminalité) qui ne sont pas directement liés à la pandémie, et une baisse du nombre de tweets faisant référence à des raisons de santé.

LES POLITIQUES DES PLATEFORMES SUR LES DISCOURS DE HAINE PENDANT LA PANDÉMIE

« Je crois fermement que Facebook ne devrait pas être l'arbitre de la vérité de tout ce que les gens disent en ligne ».

Cette phrase, prononcée par Mark Zukerberg à plusieurs reprises au fil des ans, résume bien l'attitude des plateformes en matière de modération des contenus jusqu'en 2020. Même après les élections américaines de 2016, Facebook, Twitter et YouTube ont fait face à de sérieuses critiques pour leur rôle dans la diffusion de la désinformation, de la haine et des théories du complot, mais sont restés très réticents à prendre des mesures.

En 2020, cela a changé. Facebook, Twitter et YouTube ont apporté à leurs normes communautaires et à leurs conditions d'utilisation des modifications qu'ils avaient été réticents à faire pendant des années, qu'il s'agisse de qualifier de fausses informations les comptes de personnes publiques ou de supprimer les publications d'un président américain en exercice et de supprimer son compte.

En juin 2020, la mort de l'Afro-Américain George Floyd à la suite de son arrestation par quatre policiers de Minneapolis a déclenché une vague de protestations mondiales contre le racisme et la brutalité policière. Le président américain de l'époque, Donald Trump, a publié **une série de messages sur ses plateformes de médias sociaux et dans l'un d'eux en particulier, il a écrit : « Quand le pillage commence , la fusillade commence ».** Cela a été interprété par une grande partie de la communauté afro-américaine comme une menace pour les manifestants. Twitter a décidé de cacher le contenu. Facebook ne l'a pas fait.

Au milieu des critiques, le PDG de Facebook, Mark Zuckerberg, a écrit un message dans lequel il explique les raisons pour maintenir la publication du message de Trump. « Je ne suis pas du tout d'accord avec ce que le président a dit à ce sujet, mais je pense que les gens devraient le voir par eux-mêmes, parce qu'en fin de compte, la responsabilité de ceux qui occupent des postes de pouvoir ne peut être assumée que lorsque leur discours est ouvertement examiné », a-t-il écrit.

Quelques semaines plus tard, un groupe d'entreprises - dont Unilever, Coca Cola, Verizon et Honda - a annoncé le lancement de la campagne «Stop Hate for Profit» et la suspension pendant un mois de l'achat de publicités sur la plateforme. Le vice-président d'Unilever chargé des médias, Luis Di Como, a déclaré que le fait de continuer à faire de la publicité « sur ces plateformes à l'heure actuelle n'apporterait aucune valeur ajoutée ni aux personnes ni à la société ». **« Compte tenu de la polarisation et des élections actuelles aux États-Unis, il est nécessaire d'appliquer beaucoup plus de mesures de respect des règles dans le domaine des discours de haine », a-t-il déclaré.**

« Nous respectons profondément la décision de n'importe quelle marque et restons concentrés sur le travail important consistant à supprimer les discours haineux et à fournir des informations essentielles sur le vote », a répondu Carolyn Everson, vice-présidente du groupe commercial mondial de Facebook, lundi. « Nos conversations avec les entreprises et les organisations de défense des droits civiques portent sur la manière dont nous pouvons être une force du bien ensemble. »

Toutefois, en janvier 2021, Donald Trump a été suspendu pour une durée indéterminée de Twitter et de Facebook et certaines de ses vidéos ont été retirées de YouTube pour avoir diffusé des messages dénonçant des fraudes électorales présumées lors des dernières élections américaines à l'intention de certains de ses partisans qui ont pris d'assaut le Capitole à Washington, générant de forts épisodes de violence et de peur parmi les législateurs et les fonctionnaires, ainsi que **la mort de personnes.**

« Les événements choquants des dernières 24 heures démontrent clairement que le président Donald Trump a l'intention d'utiliser le temps qu'il lui reste à exercer ses fonctions pour compromettre la transition pacifique et légale du pouvoir à son successeur élu, Joe Biden », a écrit. Zuckerberg dans un post Facebook expliquant la décision de blocage.

FACEBOOK ET LA SUPPRESSION DES CONTENUS HAINEUX

Dans ses normes communautaires, Facebook définit spécifiquement le discours de haine comme « une attaque directe contre des personnes en fonction de ce que nous appelons des « caractéristiques protégées »: race, ethnicité, origine nationale, handicap, religion, classe, orientation sexuelle, sexe, identité de genre et maladie grave ».

« Nous définissons une attaque comme un langage violent ou déshumanisant, des stéréotypes nuisibles, des affirmations d'infériorité, des expressions de mépris, de répulsion ou de rejet, des insultes ou des incitations à l'exclusion ou à la ségrégation. Nous considérons l'âge comme une caractéristique protégée lorsqu'il est mentionné avec une autre caractéristique protégée. Nous protégeons également les réfugiés, les migrants, les immigrants et les demandeurs d'asile contre les attaques graves, bien que nous autorisions les commentaires et les critiques liés aux politiques d'immigration. De même, nous offrons certaines protections pour des caractéristiques, telles que la profession, lorsqu'elles sont mentionnées en conjonction avec une caractéristique protégée », explique Facebook.

Il ajoute : « Nous sommes conscients que, parfois, les gens partagent des contenus comprenant un langage d'incitation à la haine prononcé par une autre personne dans l'intention de le désapprouver ou de sensibiliser les autres. Dans d'autres cas, un langage qui, autrement, violerait nos normes, peut être utilisé de manière autoréférentielle ou motivée. Nos politiques sont conçues pour permettre ce type de langage, mais nous exigeons que l'intention soit claire. Si tel n'est pas le cas, le contenu peut être supprimé. »

L'entreprise classe les discours de haine en trois niveaux, en fonction de la gravité de ce qui est publié sur le réseau social. Le niveau 1 correspond à tout « contenu dirigé contre une personne ou un groupe de personnes présentant des « caractéristiques protégées», qui comprend « un langage incitant à la violence ou la soutenant, que ce soit sous forme écrite ou visuelle » et « un langage ou des images déshumanisants sous forme de comparaisons, de généralisations ou de déclarations fondées sur un comportement inapproprié (sous forme écrite ou visuelle) en relation avec : les insectes, les animaux culturellement perçus comme intellectuellement ou physiquement inférieurs, la saleté, les bactéries, les maladies et les excréments, les prédateurs sexuels, la sous-humanité, les criminels sexuels et violents, les autres criminels (y compris, par exemple, les « voleurs », les « bandits »), les déclarations niant l'existence, les moqueries sur le concept de crimes de haine, les événements de crimes de haine ou leurs victimes, même si une personne réelle n'apparaît pas dans une image ».

Sont également considérés comme des discours de haine de niveau 1 « certaines comparaisons, généralisations ou déclarations fondées sur des comportements déshumanisants (écrites ou visuelles) qui incluent les Noirs et les singes ou les créatures ressemblant à des singes, les Noirs et les machines agricoles, les caricatures de Noirs avec le visage peint en noir, les Juifs et les rats, les Juifs contrôlant le monde ou des institutions importantes telles que les réseaux de médias, l'économie ou le gouvernement, la négation ou la déformation des informations sur l'Holocauste, les Musulmans et les porcs, les Musulmans et les relations sexuelles avec des chèvres ou des porcs, les Mexicains et les créatures ressemblant à des vers, les femmes en tant qu'objets domestiques ou la référence aux femmes en tant que propriété ou 'objets'», ainsi que « la référence aux personnes transgenres ou de genre non binaire comme si elles n'étaient pas des êtres humains ou aux Dalits ou aux personnes de castes registrées ou 'basses' en tant que domestiques ».

Les discours de haine de niveau 2 pour Facebook sont ceux qui font référence à des groupes protégés et comprennent des « généralisations dénotant une infériorité (sous forme écrite et visuelle) » telles que des « déficiences physiques » liées à « l'hygiène, y compris, mais sans s'y limiter : 'crasseux', 'sale'», à « l'apparence physique telles que 'laid', 'hideux' », aux « déficiences mentales» telles que 'débile', 'stupide', 'idiot', à « l'éducation » telles que 'analphabète', 'sans éducation', à la « santé mentale » telles que 'malade mental', 'retardé', 'fou', 'insensé' et aux « déficiences morales » liées aux « traits de personnalité qui sont considérés comme culturellement

négatifs, y compris, entre autres : «lâche», «menteur», «arrogant», «ignorant»» et « des termes péjoratifs liés à l'activité sexuelle » tels que «salope», «putain», «traînée», «pervers».

Sont également inclus dans le niveau 2 des « expressions dénotant une insuffisance » telles que «inutile», «nul», des « expressions de supériorité ou d'infériorité par rapport à une autre caractéristique protégée », des « expressions liées à un écart par rapport à la norme » telles que «anormal» et des « expressions de mépris » telles que la «reconnaissance de l'intolérance à l'égard de caractéristiques protégées » telles que «homophobe», «islamophobe», « raciste », ainsi que des « expressions indiquant qu'une caractéristique protégée ne devrait pas exister » et des « expressions de haine », «rejet » et « répulsion » telles que «haine» ou «aucun respect», «n'aime pas», «s'en fout», «vomit», «dégoûtant», «désagréable», etc. Sont également incluses dans cette catégorie les insultes « liées aux organes génitaux ou à l'anus pour désigner une personne », les « phrases ou termes offensants destinés à insulter », les « termes ou phrases qui incitent à se livrer à des activités sexuelles ou qui font référence au contact avec les organes génitaux ou l'anus, ou avec les fèces ou l'urine ».

Enfin, Facebook classe comme discours de haine au niveau 3 les contenus en image ou en texte qui font référence à la « ségrégation sous forme d'incitation, de déclarations d'intention, de plaidoyer ou de soutien, ou de déclarations d'aspirations ou de conditions en rapport avec la ségrégation », à « l'exclusion sous forme d'incitation, de déclarations d'intention, de plaidoyer ou de soutien, ou de déclarations d'aspirations ou de conditions qui incluent une exclusion explicite, c'est-à-dire des actes tels que l'expulsion de certains groupes ou l'indication qu'ils ne sont pas autorisés », l'exclusion politique, c'est-à-dire le déni du droit à la participation politique, l'exclusion économique, c'est-à-dire le déni de l'accès aux avantages financiers et la limitation de la participation au marché du travail, l'exclusion sociale, c'est-à-dire des actes tels que le déni de l'accès à certains espaces (physiques et en ligne) et services sociaux » et « les contenus qui décrivent ou singularisent négativement des individus par la stigmatisation, la stigmatisation étant définie comme des mots intrinsèquement offensants utilisés comme des étiquettes insultantes ».

En juillet 2020, à la suite de **la campagne Stop Hate for Profit** plus de 1 200 entreprises du monde entier se sont jointes à un boycott publicitaire contre les grandes plateformes, appelant à une modération accrue des discours haineux ainsi qu'à la suspension de la publicité des comptes qui encouragent la discrimination contre des groupes spécifiques. L'une des principales demandes des organisations et entreprises concernées était la suppression de tous les comptes de Trump.

La coalition demande notamment aux plateformes de supprimer « les groupes ou les pages axés sur la suprématie blanche, les milices, l'antisémitisme, l'islamophobie et les conspirations violentes », d'«augmenter les ressources consacrées à la surveillance des discours de haine et des groupes violents », de « modifier la politique des plateformes afin d'interdire toute page d'événement qui appelle aux armes à feu», ainsi que « s'engager à consacrer 5 % de ses revenus annuels au financement d'un fonds indépendant destiné à soutenir les initiatives, universitaires et d'organisations, qui luttent contre le racisme, la haine et la division causée par l'inaction de Facebook ».

Suite à ces réclamations, Facebook a publié un post en juin 2020 **dans lequel il répondait à certaines des demandes de Stop Hate for Profit**. En ce qui concerne la demande de l'organisation de « créer une modération séparée composée d'experts en haine fondée sur l'identité pour les utilisateurs qui expriment qu'ils ont été attaqués », Facebook a assuré que « les signalements de discours haineux sur Facebook sont déjà automatiquement acheminés vers un ensemble de réviseurs ayant reçu une formation spécifique sur nos politiques de haine fondée sur l'identité dans 50 marchés couvrant 30 langues » et qu'en outre, il y a « des consultations avec des experts en haine fondée sur l'identité pour développer et faire évoluer les politiques que ces réviseurs formés appliquent ». Ils ont également annoncé leur « intention d'inclure la prévalence du discours de haine dans les futurs rapports de conformité aux normes communautaires (CSER), en attendant qu'il n'y ait plus de complications liées à la COVID-19 ».

Le même mois, le vice-président des politiques publiques de Facebook, Richard Allan, a écrit une chronique dans laquelle il a abordé les différences dans la définition du discours de haine dans différentes parties du monde et les difficultés de la plateforme à les **détecter correctement et à prendre les mesures correspondantes**. « Il n'y a pas de réponse universellement acceptée lorsque quelqu'un franchit la ligne. Même si certains pays disposent de lois contre les discours de haine, leurs définitions varient considérablement. En Allemagne, par exemple, les lois interdisent l'incitation à la haine ; quelqu'un pourrait faire l'objet d'une descente de police pour avoir mis en ligne un tel contenu. Aux États-Unis, en revanche, même les discours les plus vils sont légalement protégés par la Constitution américaine », écrit M. Allen. «Des personnes vivant dans le même pays - ou à côté - ont souvent des niveaux de tolérance différents aux discours. Pour certains, un humour caustique sur un chef religieux peut être considéré comme un blasphème et un discours de haine contre tous les adeptes de cette foi. Pour d'autres, une bataille basée sur des insultes sexistes peut être un moyen mutuellement agréable de partager un rire. Est-il acceptable qu'une personne publie des choses négatives sur des personnes d'une certaine nationalité si elle est de la même nationalité ? Que se passe-t-il si un jeune qui fait référence à un groupe ethnique particulier en utilisant des insultes raciales cite les paroles d'une chanson ? » le dirigeant de Facebook se demande.

Allen fait également référence dans le texte aux erreurs commises lors de la suppression de contenus classés à tort comme discours de haine. « Si nous ne supprimons pas le contenu que vous signalez comme étant un discours de haine, nous estimons que nous ne respectons pas les valeurs de nos normes communautaires. Lorsque nous supprimons un message que vous publiez et que vous croyez qui représente un point de vue raisonnable, cela peut être ressenti comme une censure. Nous savons à quel point les gens se sentent mal lorsque nous commettons ces erreurs, c'est pourquoi nous nous efforçons constamment d'améliorer nos processus et d'expliquer davantage les choses », explique le dirigeant de Facebook.

Il ajoute que les erreurs de Facebook en matière de modération de contenu « ont

suscité une grande inquiétude dans certaines communautés, notamment les groupes qui estiment que nous agissons - ou n'agissons pas - en raison de nos préjugés ». « L'année dernière (2019), Shaun King, un éminent militant afro-américain, a posté un courriel d'incitation à la haine qu'il avait reçu et qui comprenait des insultes racistes. Nous avons supprimé le post de King par erreur, car nous n'avons pas reconnu initialement qu'il était partagé pour condamner l'attaque», a-t-il déclaré. En juillet, Nick Clegg, vice-président des affaires et des communications mondiales de Facebook, a écrit **un article dans lequel il affirmait que l'entreprise avait pris plusieurs mesures qui avaient permis de réaliser des progrès significatifs dans l'élimination des discours de haine sur sa plateforme**. « Un rapport récent de la Commission européenne a révélé que Facebook évaluait 95,7 % des signalements de discours haineux en moins de 24 heures, plus rapidement que YouTube et Twitter », a écrit Clegg. « Le mois dernier, nous avons indiqué que nous avons trouvé près de 90 % des discours haineux que nous avons supprimés avant que quelqu'un ne les signale, contre 24 % il y a un peu plus de deux ans. Nous avons pris des mesures contre 9,6 millions de contenus au premier trimestre 2020, contre 5,7 millions au trimestre précédent. Et 99 % du contenu d'ISIS et d'Al-Qaïda que nous supprimons est éliminé avant que quiconque nous le signale », a-t-il ajouté.

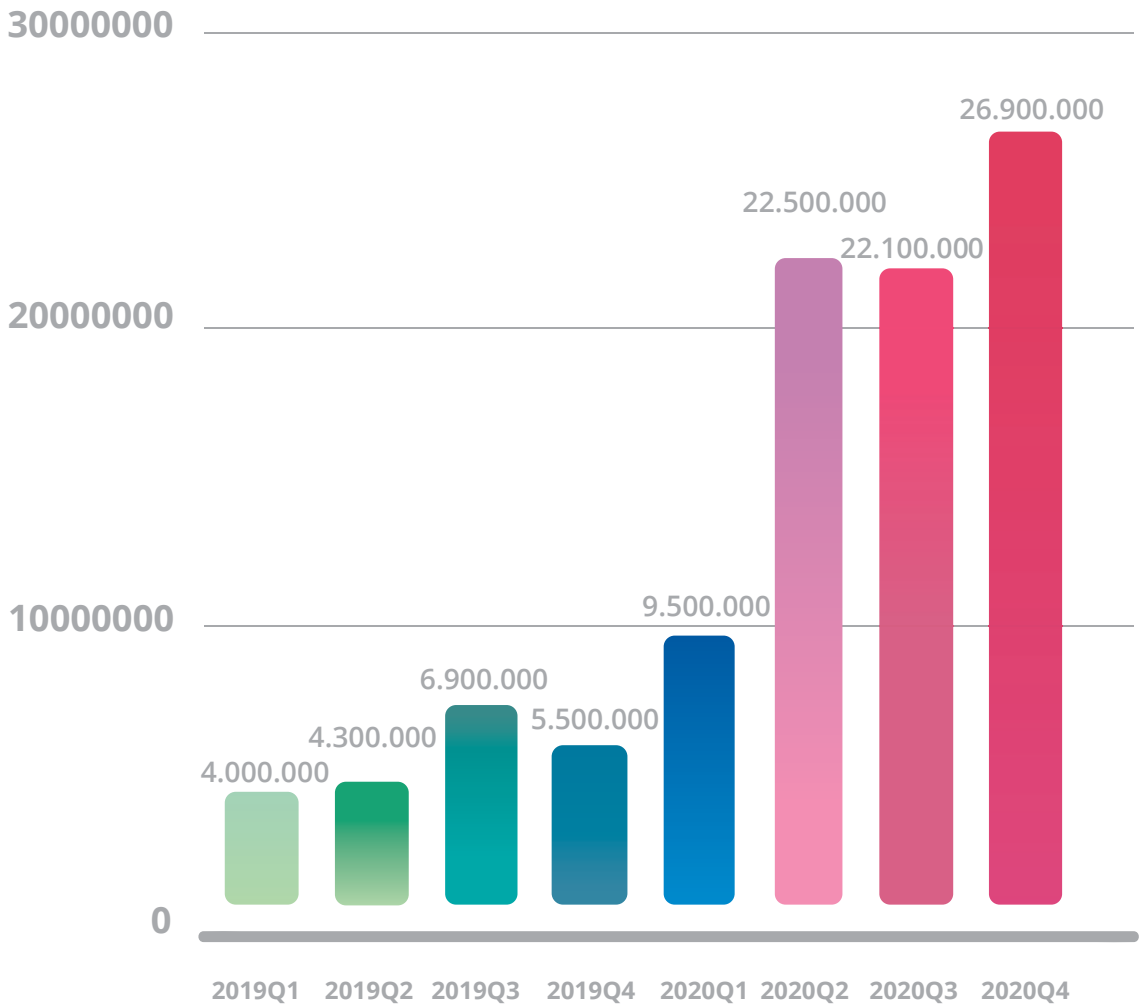
Selon le Community Standards Enforcement Report (CSER) publié en février 2021, le nombre de contenus sur lesquels Facebook a pris des mesures est passé de 20 700 000 en 2019 à 81 000 000, soit une augmentation de près de 300 % de la quantité de contenus catégorisés comme discours de haine entre une année et l'autre.

En novembre, Facebook a commencé à mesurer la prévalence des discours de haine sur le réseau social et a constaté qu'entre juillet et septembre, ce chiffre se situait entre 0,10 % et 0,11 %. Cela signifie que sur 10 000 consultations de messages sur le réseau social, entre 10 et 11 seraient classées comme discours de haine, selon Facebook. Entre octobre et décembre 2020, ce chiffre est tombé à 0,07 % à 0,08 % de prévalence. Dans son rapport, Facebook ne précise pas si ce changement est dû à une augmentation du nombre total de messages, à une diminution réelle de la catégorie des discours haineux par rapport au total, ou à un changement des critères ou processus de détection.

Si l'on examine l'année 2020 en profondeur, on peut détecter une augmentation très significative du nombre de contenus de discours de haine sur lesquels Facebook a pris des mesures à partir du deuxième trimestre de 2020. Entre janvier et mars, des mesures ont été prises par rapport à 9 500 000, tandis que les mois suivants, le chiffre a doublé pour atteindre 22 500 000 entre avril et juin, 22 100 000 entre juillet et septembre, et 26 900 000 entre octobre et décembre.

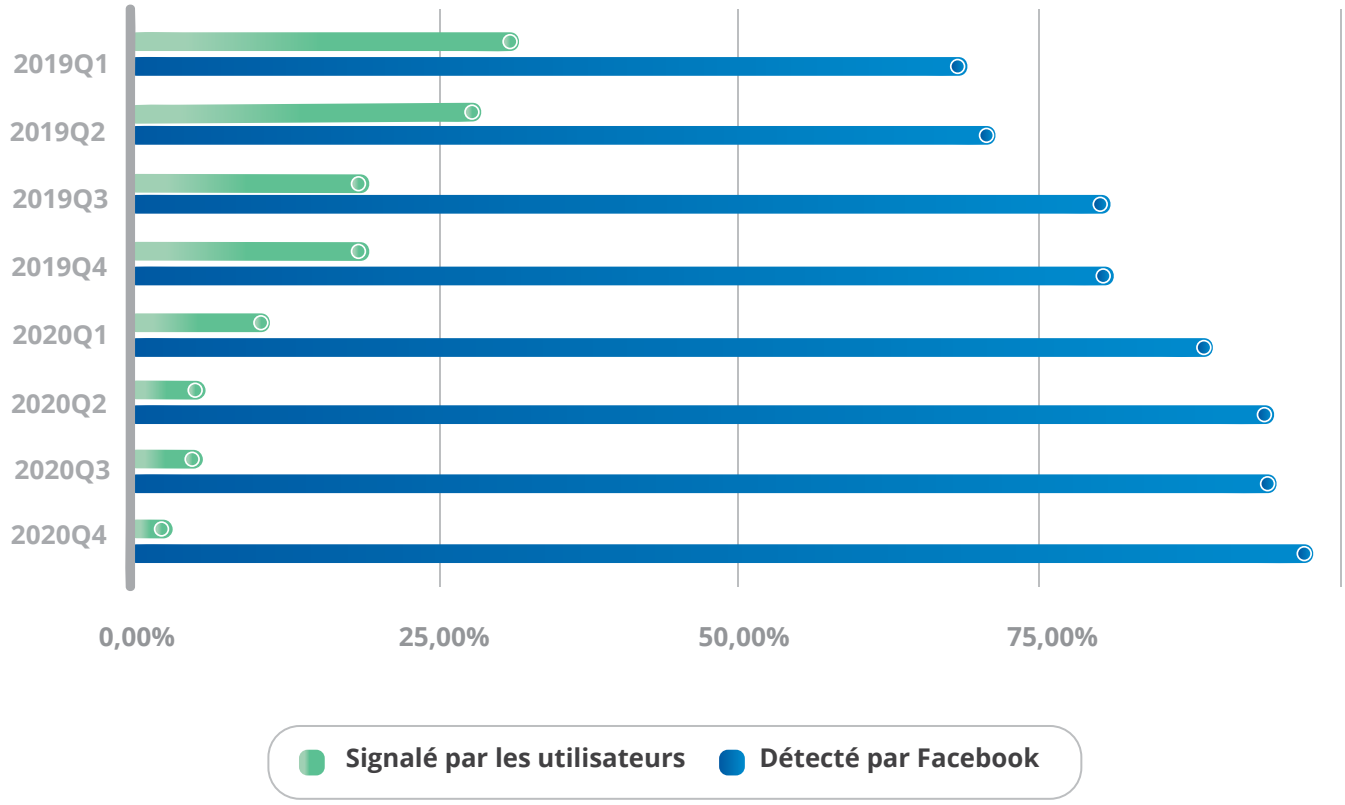
Selon le rapport du CSER, la croissance du nombre de posts détectées ainsi que du pourcentage de détection proactive est «principalement due à l'amélioration des systèmes de technologie de détection en arabe et en espagnol » ainsi qu'à « l'expansion de l'automatisation en portugais ».

Contenu sur lequel des actions ont été prises par rapport aux discours de haine



Un autre aspect frappant est l'augmentation du pourcentage de contenu détecté par Facebook par rapport à celui signalé par les utilisateurs sur le total des contenus ayant fait l'objet d'une action en raison de la catégorie des discours de haine. Avec de petites fluctuations, le poids des systèmes internes de Facebook sur la quantité totale de contenus haineux augmente régulièrement depuis 2018, pour atteindre presque le total au dernier trimestre de 2020. Au dernier trimestre 2017, Facebook a agi sur 1 700 000 de discours de haine, dont 76,4 % ont été détectées à partir de rapports d'utilisateurs. En 2020, ce rapport s'est inversé. Entre janvier et mars, 89,3 % des contenus classés comme discours haineux provenaient des systèmes de détection de Facebook. La situation était similaire entre avril et juin (94,7 %), juillet et septembre (94,7 %) et octobre et décembre (97,1 %).

Combien a été dénoncé du contenu “Discours de haine”



Quelque chose de similaire s’est produit avec Instagram, également détenu par Facebook, où l’action sur les contenus haineux est mesurée depuis le dernier trimestre de 2019. Entre janvier et mars 2020, Instagram a détecté et pris des mesures à l’égard de 578 000 contenus comme relevant de sa définition du discours de haine, puis 3 200 000 entre avril et juin, 6 500 000 entre juillet et septembre, et 6 600 000 entre octobre et décembre 2020.

Au cours du premier trimestre de l’année, 57,1 % du contenu a été détecté à partir des rapports des utilisateurs, tandis qu’au cours du trimestre suivant, le rapport a changé radicalement et les rapports des utilisateurs n’ont représenté que 15,1 % du total des actions entreprises sur les discours haineux. Ce ratio est resté le même au cours des trimestres suivants : 5,2 % entre juillet et septembre, 4,9 % entre octobre et décembre.

À la mi-mars 2020, et après des demandes continues de ces équipes suite aux mesures d’isolement à cause de la pandémie de COVID-19, Facebook a décidé d’envoyer ses plusde15000modérateursdecontenurépartis sur 20 sites différents faire du télétravail.

Mark Zuckerberg, PDG de Facebook, a déclaré cette semaine-là que Facebook serait contraint, pendant la pandémie qui balaie la majeure partie du monde, de « s'appuyer plus activement sur des logiciels d'intelligence artificielle pour prendre des décisions de modération de contenu ». La société a également déclaré qu'elle organiserait une formation à plein temps pour que ses employés faisaient « davantage attention » aux contenus « hautement sensibles ». Elle a prévenu que les utilisateurs « devraient s'attendre à davantage d'erreurs à mesure que Facebook améliorait le processus, en partie parce que seule une fraction d'humains y participerait encore et parce que le logiciel prend des décisions plus naïves que les humains, ce qui pourrait entraîner des « faux positifs », y compris la suppression de contenus qui n'auraient pas dû l'être. « Cela créera un compromis contre certains types de contenus qui ne présentent pas de risques physiques imminents pour les gens », **a déclaré Zuckerbeerg.**

En novembre 2020, Facebook **a annoncé des changements dans ses systèmes de modération qui impliquent** une présence accrue de la modération automatisée dans les premières étapes de contact avec le contenu. Chris Palow, ingénieur et membre de l'équipe « Intégrité » de Facebook, a admis lors de la conférence de presse que « l'intelligence artificielle ne sera jamais parfaite » et qu'elle « a ses limites » lorsqu'il s'agit de séparer les discours de haine de ceux qui ne le sont pas, par exemple, quand il s'agit d'une parodie ou de l'humour. « Le système vise à combiner l'intelligence artificielle et l'examen humain afin de réduire le nombre d'erreurs », a-t-il déclaré. Facebook ne rend pas public le pourcentage de contenu qui est classé à tort comme contenu à supprimer.

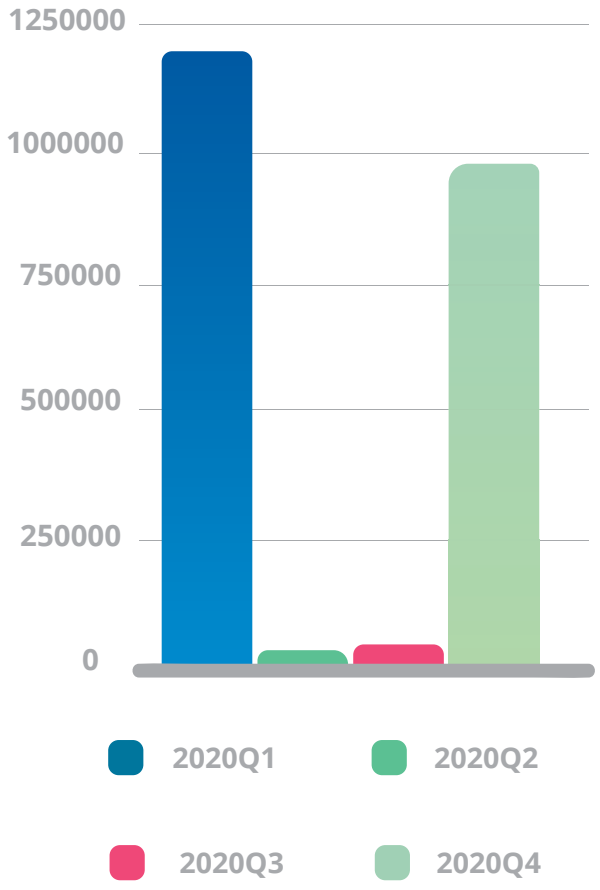
Des mois plus tard, en février 2021, le responsable de la politique de contenu organique de Facebook, Varun Reddy, a déclaré que la plateforme rencontrait des problèmes liés à l'absence de modérateurs humains dans le processus de modération d'une grande partie de son contenu. L'intelligence artificielle apprend des modérateurs humains, a-t-il expliqué, et la réduction de la présence humaine a modifié **« l'efficacité de l'intelligence artificielle au fil du temps ».**

« Nous travaillons avec les fournisseurs pour avoir le plus de capacité de remettre en ligne (...). Nous n'en sommes pas encore au point de départ, mais depuis le début du confinement, le 25 mars, nous n'en sommes plus là. Nous espérons que dans les mois à venir les systèmes retrouveront leur pleine efficacité », a déclaré M. Reddy en février de cette année.

Un autre aspect touché par l'isolement des employés de Facebook est la procédure d'appel pour les contenus que les utilisateurs estiment avoir été injustement supprimés. « En raison d'une réduction temporaire de notre capacité d'examen à la suite de la COVID-19, nous ne pouvons pas toujours offrir aux utilisateurs la possibilité de faire appel. Nous avons donné aux gens la possibilité de nous dire qu'ils n'étaient pas d'accord avec notre décision, ce qui nous a aidés à réexaminer un grand nombre de ces cas et à restaurer le contenu dans les cas où cela

était approprié », indique Facebook dans son **rapport CSER**. On peut y voir qu'entre avril et juin 2020, les appels ont été quasi inexistants, n'atteignant que 70 000 dans le monde entier au cours de ces 6 mois, alors qu'au trimestre précédent, ils avaient atteint 1 200 000. Dans la période suivante, entre octobre et décembre, les appels ont atteint 984 200 cas.

Appels sur le contenu ayant fait l'objet d'une action



En 2020, Facebook a également atteint un nombre record de contenus restaurés par rapport aux périodes précédentes, passant de 483 400 contenus en 2019 à 703 200 en 2020. Parmi ces derniers, Facebook en a rétabli 589 300 sans n'avoir reçu aucun appel.

LA MODÉRATION DES DISCOURS DE HAINE SUR

En décembre 2020, Twitter a annoncé une mise à jour de ses règles pour lutter contre la propagation des discours de haine sur sa plateforme et a étayé sa décision par « des recherches liant le langage déshumanisant à la violence hors ligne ». En 2019, Twitter a mis à jour ses règles concernant les discours de haine pour inclure la religion et la classe comme groupes protégés, en mars 2020 ils ont ajouté l'âge, le handicap et la maladie, et en décembre 2020 **ils ont annoncé l'interdiction des propos qui déshumanisent les personnes en fonction de leur race, de leur ethnie ou de leur nationalité.**

La publication comprenait un certain nombre d'exemples pour illustrer les discours qui ne seraient pas autorisés après l'annonce :

> ***« Tous les (nationalité) sont des cafards qui vivent des allocations de l'État et devraient être expulsés », « Les gens qui sont (race) sont des sangsues et ne sont bons qu'à une seule chose », « Il y a trop de (nationalité, race, ethnie) fumiers dans notre pays et ils devraient partir », « Tous les (groupe d'âge) sont des sangsues et ne méritent pas notre soutien », « Les gens qui ont (maladie) sont des rats qui polluent tout ce qui les entoure », « (Groupe religieux) devrait être puni. Nous ne faisons pas assez pour nous débarrasser de ces animaux puants ».***

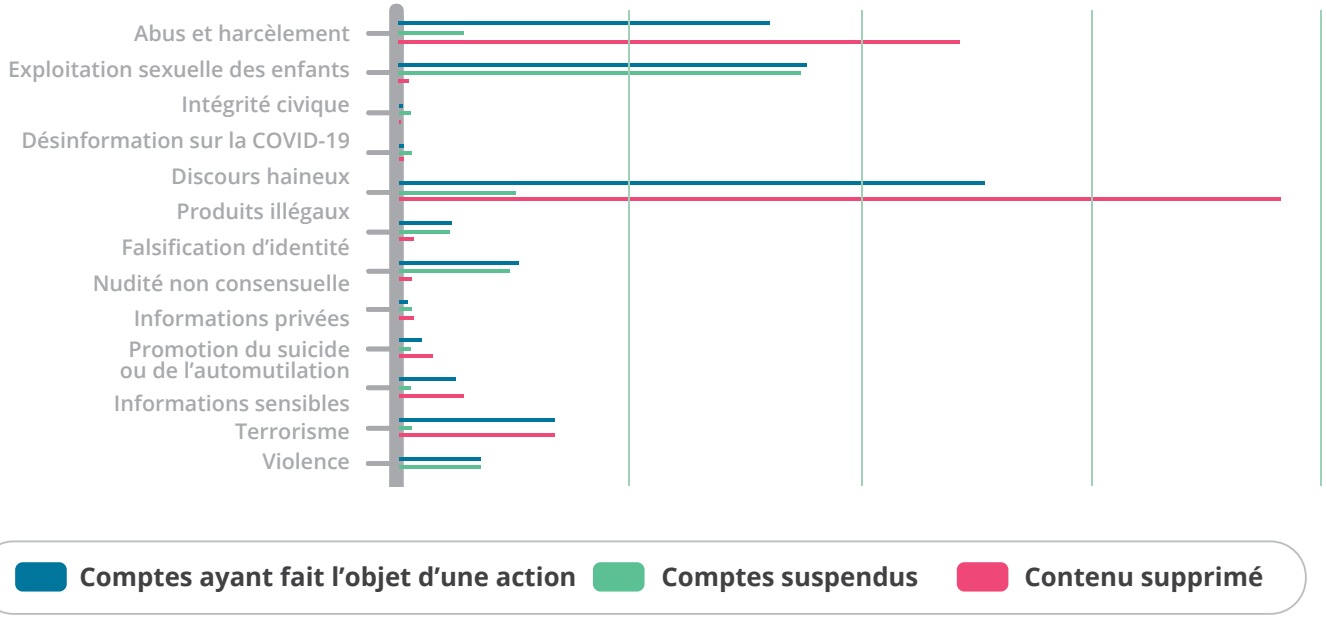
En octobre 2019, Kamala Harris, désormais vice-présidente des États-Unis, a publié **une lettre ouverte** à Jack Dorsey, PDG de Twitter, dans laquelle elle demandait la modération de certains messages de Donald Trump, alors président, car elle estimait qu'ils violaient les normes communautaires du réseau social, notamment celles relatives aux discours de haine. « Aucun utilisateur, quel que soit son emploi, sa richesse ou son statut, ne devrait être exempté de respecter les règles d'utilisation de Twitter », a déclaré M. Harris dans sa lettre.

La même année, **une étude de l'université de New York (NYU)** a montré une corrélation entre le nombre de tweets racistes et le nombre de crimes de haine racistes dans 100 villes des États-Unis. « Je pense qu'il y a un sentiment dans les tweets trouvés qui est lié au fait de favoriser un environnement propice à ces crimes », explique Rumi Chunara, l'un des auteurs de l'étude. Il ajoute qu'à l'inverse, **« avoir des conversations productives améliore l'environnement et les résultats ».**

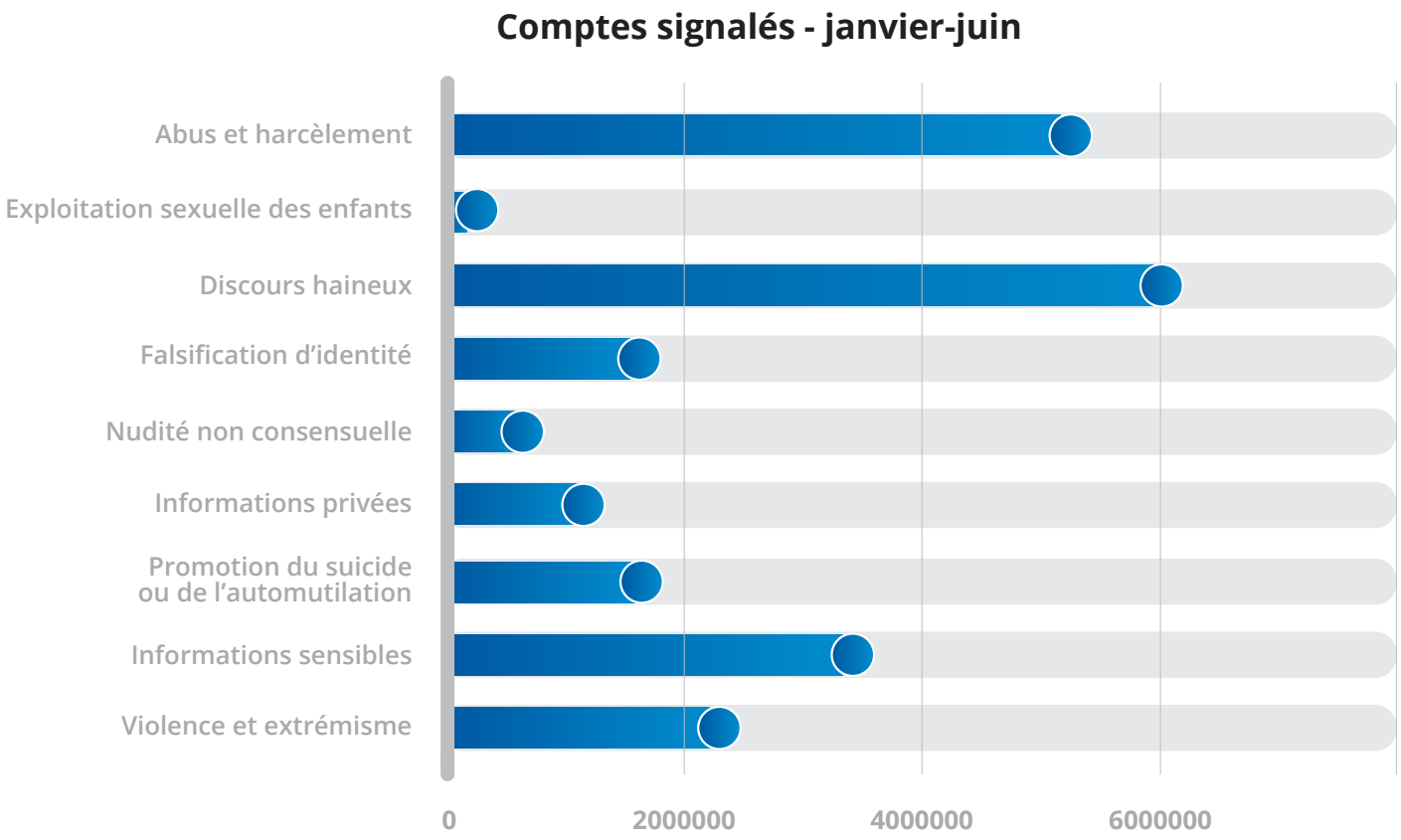
« Actuellement, le système rend super facile d'intimider et d'abuser des autres», a déclaré Dorsey en 2019, ajoutant que «l'un des problèmes est la taille du poids qu'il accorde aux followers et aux likes ».

Que s'est-il passé en 2020 et pendant la pandémie de COVID-19 ? Selon **le dernier rapport disponible Twitter** de Twitter Transparency Report, entre janvier et juin de cette année-là, des mesures ont été prises à l'égard de 1 940 082 comptes, dont 925 954 ont été suspendus et 1 927 063 contenus ont été supprimés. Un nombre très similaire de contenus a été supprimé au cours de la même période en 2019 (1 914 471), mais moins de comptes ont été suspendus (687 397).

Comptes ayant fait l'objet d'une action - janvier-juin - pour des raisons de



Sur ce total, 645 416 comptes ont fait l'objet d'une action en raison d'un contenu qualifié de discours de haine, soit 33,2 % des comptes ayant fait l'objet d'une action. 12 400 000 comptes ont été signalés au cours de la période janvier-juin, dont près de la moitié (6 055 642) pour des raisons de discours de haine. Selon le rapport, 30 % de comptes supplémentaires ont été signalés par rapport à la même période l'année dernière.



Twitter fait état d'une réduction de 35 % des comptes ayant fait l'objet d'une action pour discours haineux par rapport à la période précédente, bien qu'il reconnaisse que, dans ces circonstances, les équipes se sont concentrées sur l'examen des contenus susceptibles de causer des dommages ou liés à des informations erronées sur la COVID-19 et ont enregistré « des retards importants dans tous les autres domaines ».

En avril 2020, Twitter a publié un billet de blog informant de certains changements résultant de la décision d'envoyer un grand nombre de ses employés chez eux pour observer les mesures de distanciation sociale poussées par les gouvernements du monde entier.

Une partie de ces mesures consistait en «une utilisation accrue de l'apprentissage automatique et de l'automatisation pour prendre un large éventail de mesures sur les contenus potentiellement abusifs et manipulateurs ». « Nous voulons être clairs: bien que nous efforcions de rendre les systèmes cohérents, il peut arriver que le manque de contexte nous fasse commettre des erreurs. Par conséquent, nous ne procéderons pas à une suspension permanente sur la base des seuls systèmes de modération automatisés. Au lieu de cela, nous continuerons à rechercher les occasions où les contrôles de modération humaine ont le plus d'impact », indique le texte.

Twitter a indiqué que pendant la pandémie de COVID-19, une technologie automatisée serait utilisée pour « attirer l'attention sur les contenus les plus susceptibles de causer des dommages et qui seront examinés en premier» et pour « identifier de manière proactive les violations des règles avant qu'elles ne soient signalées (les équipes apprennent sur la base des décisions passées, de sorte qu'avec le temps, la technologie peut aider à classer les contenus ou à examiner les comptes automatiquement) ». Pour les contenus qui « nécessitent un contexte supplémentaire, comme des informations trompeuses sur la COVID-19 », Twitter indique que ses équipes « continueront à examiner manuellement les signalements ».

Le réseau social précise que les délais de réponse aux signalements s'étendront « au-delà des délais normaux » et admet que « étant donné que les systèmes automatisés n'ont pas tout le contexte et la perspicacité des équipes humaines, des erreurs seront commises ».

YOUTUBE ET LA MODÉRATION DES DISCOURS DE HAINE EN TEMPS DE PANDÉMIE

Sur YouTube, la dernière mise à jour des normes communautaires concernant les discours de haine date de 2019. Actuellement, la définition du discours de haine de l'entreprise appartenant à Google est la suivante : « contenu qui encourage la violence et la haine à l'encontre d'individus ou de groupes en fonction de l'un des attributs suivants : âge, caste, handicap, ethnie, identité de genre, nationalité, race, statut migratoire, religion, sexe ou genre, orientation sexuelle, victimes d'un événement violent ou leurs familles, et anciens combattants ».

Les normes communautaires ajoutent que YouTube ne permet pas de « déshumaniser les individus ou les groupes présentant ces caractéristiques, en affirmant qu'ils sont physiquement ou mentalement inférieurs, ou en faisant l'éloge ou la glorification de la violence à leur encontre », ni « d'utiliser des stéréotypes qui incitent ou encouragent la haine fondée sur ces caractéristiques, ou des insultes raciales, des insultes ethniques, religieuses ou autres dont le but premier est de promouvoir la haine », qui « revendique la supériorité d'un groupe sur ceux qui présentent l'une des caractéristiques ci-dessus pour justifier la violence, la discrimination, la ségrégation ou l'exclusion» ou qui « nie que des événements violents bien documentés se sont produits ».

En mars 2021, YouTube a fait l'objet d'un débat animé sur ses politiques de modération des discours haineux lorsqu'il a supprimé une vidéo du commentateur Steve Crowder pour considérer qu'elle violait ses politiques relatives à la diffusion de fausses informations sur la COVID-19. Dans cette vidéo, M. Crowder a fait une série de commentaires sur la décision du gouvernement républicain d'accorder une subvention aux agriculteurs des minorités raciales au motif qu'ils ont été historiquement exclus des politiques d'aide à ce secteur. **Les commentaires comprenaient des caractérisations de la façon dont les Afro-Américains parlent, bougent et pensent.**

Après des plaintes de diverses organisations de défense des droits des minorités raciales, YouTube a publié une déclaration dans laquelle il assure que ses « politiques interdisent les contenus qui encouragent la haine envers des groupes en fonction de leur race » mais que « bien qu'offensante, cette vidéo de Steven Crowder ne viole pas ces politiques ».

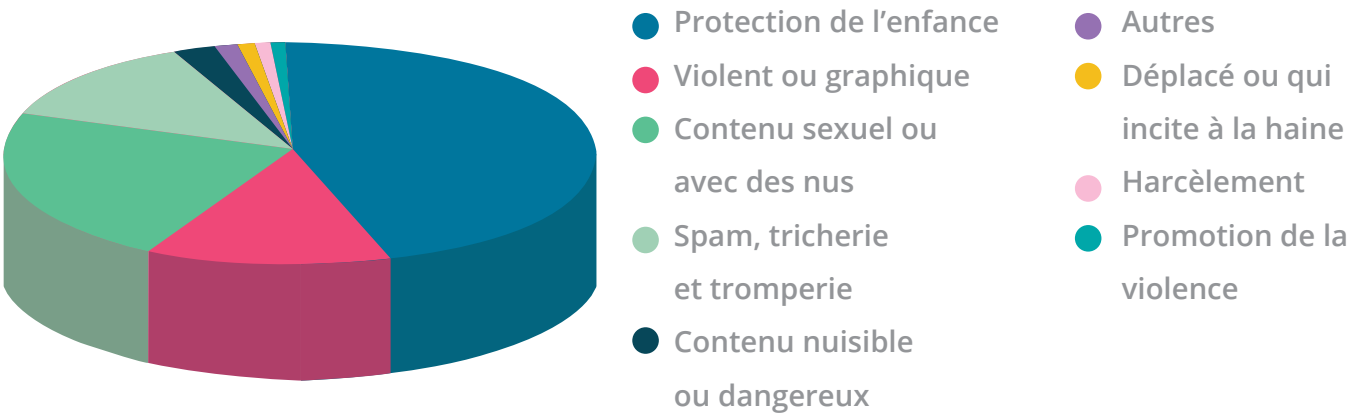
En avril 2021, YouTube a publié des informations affirmant qu'il avait amélioré ses systèmes de détection des discours haineux sur la plateforme. « Nous ne voulons pas que YouTube soit une plateforme qui puisse causer du tort dans le monde de manière flagrante », **a déclaré le chef de produit de la plateforme, Neal Mohan.**

Sur YouTube, le phénomène semble plus complexe à détecter. En fait, les données disponibles ne permettent pas d'affirmer qu'il y a eu une augmentation significative des discours de haine sur la plateforme, bien que des épisodes isolés aient été mis en évidence par les médias et l'opinion publique.

Entre avril et juin 2020, YouTube a supprimé 11 401 696 vidéos, sans compter plus de 30 000 000 millions de vidéos qui ont été supprimées à la suite de la suppression de 1 998 635 chaînes au cours de la même période. Sur ces plus de onze millions de vidéos, seules 552 062 vidéos ont été supprimées sans recourir à des systèmes de détection automatique. Entre juillet et septembre, 7 872 684 vidéos ont été supprimées et seulement 481 721 l'ont été sans détection automatique, et entre octobre et décembre, 9 321 948 vidéos ont été supprimées et seulement 521 866 l'ont été sans l'utilisation de systèmes de détection automatique pour détecter les violations des règles de YouTube.

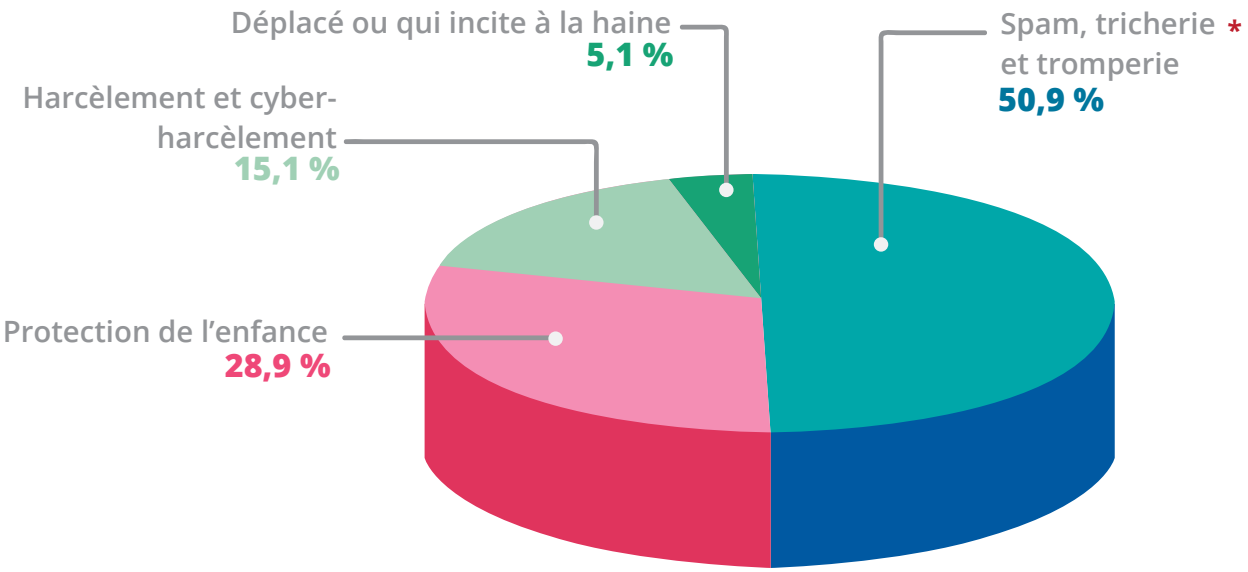
En termes de motifs, les discours de haine n'ont pas occupé un espace significatif, atteignant un chiffre de 97 362 vidéos supprimées au dernier trimestre de 2020, même s'ils ont enregistré une légère augmentation entre le trimestre avril-juin et les deux trimestres suivants, passant de 0,7 % des vidéos supprimées à plus de 1 %.

Vidéos supprimées selon la raison - octobre - décembre



Si l'on considère la suppression des commentaires sur les vidéos, soit quelque 906 196 160 supprimés au dernier trimestre de 2020, la raison « incitation à la haine » augmente à 5 % parmi les raisons spécifiées pour la suppression de ces contenus. Cela signifie qu'au cours du dernier trimestre de 2020, plus de 46 millions de commentaires ont été supprimés de YouTube parce que les systèmes de modération automatisés ont jugé qu'ils violaient les règles de la plateforme concernant la catégorisation des « discours de haine ».

Commentaires supprimés, selon le motif de suppression oct- déc



Lorsque YouTube a envoyé ses modérateurs de contenu chez eux en mars en raison de la pandémie de COVID-19 et a étendu énormément l'utilisation de ses filtres automatisés, cela a conduit à un doublement des vidéos supprimées au deuxième trimestre 2020. Cette croissance a laissé le réseau social appartenant à Alphabet ouvert au débat sur l'exactitude des processus de modération automatisés.

« En réponse à la situation de la COVID-19, nous avons pris des mesures pour protéger notre personnel externe et réduire le personnel en présentiel dans les bureaux. En conséquence, et temporairement, nous utilisons davantage la technologie pour accomplir certaines des tâches qui sont normalement effectuées par des réviseurs en chair et en os, et nous supprimons donc davantage de contenus qui pourraient ne pas enfreindre nos politiques. Cela a un impact sur certains des paramètres de ce rapport et continuera probablement à avoir un impact sur les paramètres à l'avenir », a écrit l'entreprise dans un billet de blog accompagnant **son rapport de transparence pour le dernier trimestre.** « Étant donné que la responsabilité est notre priorité absolue, nous avons choisi la seconde solution : utiliser la technologie pour aider à accomplir une partie du travail fait normalement par les réviseurs », a expliqué Google.

Dans le rapport du deuxième trimestre, YouTube a admis que l'augmentation du nombre de suppressions de contenu était due au fait que l'entreprise « a accepté un niveau d'efficacité inférieur pour être sûre de supprimer le plus grand nombre possible de contenus ». « L'une des décisions que nous avons prises au début de la pandémie, lorsqu'il est apparu clairement que les machines ne seraient pas aussi précises que les humains, a été de privilégier la protection des utilisateurs, même si cela pouvait entraîner le retrait d'un nombre légèrement plus élevé de vidéos », a déclaré Neil Mohan, chef de produit de YouTube, à **la publication spécialisée américaine Mashable.**

En septembre, YouTube a annoncé que les modérateurs humains allaient commencer à revenir aux bureaux et travailler à la refonte des systèmes de modération pour tenter de retrouver les chiffres du début de l'année 2020.

Comme indiqué plus haut, l'utilisation de systèmes de détection automatique a eu pour conséquence, selon ses propres dirigeants, que YouTube a supprimé de nombreux contenus qui ne violaient en fait pas ses normes communautaires, doublant ainsi le nombre d'appels, qui est passé de 166 000 au premier trimestre à 325 000 au deuxième trimestre 2020.

Contrairement à Facebook, YouTube n'a pas réduit l'attention portée aux processus d'appel et a continué à maintenir les délais de processus antérieurs à la COVID-19. Cela signifie que le nombre de vidéos rétablies à la suite des appels a également augmenté, passant de 41 000 à 161 000 au cours de cette période. **Cela montre que le taux de rétablissement normal de YouTube, qui est normalement de 25 % des appels, est passé à près de la moitié.**

Dans son rapport de transparence, YouTube détaille son processus spécifique de modération des discours haineux et aborde certaines des difficultés spécifiques que ce type de contenu présente par rapport à d'autres types de contenu également interdits par les normes communautaires.

« La politique en matière de discours de haine est complexe à mettre en œuvre à grande échelle, car les décisions prises nécessitent une analyse nuancée du contexte et une compréhension approfondie du langage en question. Afin d'être en mesure d'appliquer notre politique de manière cohérente, nous avons élargi notre équipe de révision par des experts sur la question et sur les sujets linguistiques. En outre, nous mettons en œuvre l'apprentissage automatique pour détecter les contenus haineux potentiels et les envoyer à l'équipe de révision, et nous appliquons les leçons apprises sur d'autres types de contenus, comme l'extrémisme violent. Parfois, nous nous trompons, c'est pourquoi nous disposons d'une procédure d'appel pour les créateurs qui estiment que leur contenu a été supprimé de manière inappropriée. Nous évaluons constamment nos politiques et nos lignes directrices en matière d'application, et nous continuerons à travailler avec des experts et la communauté pour apporter des changements le cas échéant », déclarent-ils.

YouTube ajoute qu'en plus de « supprimer le contenu » qui viole ses normes communautaires, il s'efforce de « réduire les recommandations pour le contenu qui est à la limite » de la violation de ses directives. **« Depuis un certain temps, nous avons également des directives de contenu approprié pour les annonceurs, qui interdisent de montrer des publicités sur des vidéos qui incluent des contenus haineux », disent-ils.**

CONCLUSIONS

Sous d'innombrables pressions politiques, sociales et médiatiques, Facebook, YouTube et Twitter ont, ces derniers mois, apporté des modifications à leurs normes communautaires relatives aux discours de haine et ont pris des décisions auxquelles ils semblaient fortement résister les années précédentes et qui impliquent une augmentation substantielle de leur rôle de régulateur de ce qui peut et ne peut pas être dit dans ces nouveaux espaces publics.

Il est difficile de savoir dans quelle mesure ces changements ont été couronnés de succès et même de définir le succès face à des mesures dont les plates-formes elles-mêmes admettent qu'elles ne fonctionnent pas clairement de manière adéquate pendant la pandémie de COVID-19. Ces mesures comprenaient, dans certains cas comme Facebook et Instagram, des actions très restrictives, voire injustifiées, comme la disparition virtuelle des processus d'appel pendant plusieurs mois. Cela signifiait non seulement la suppression de contenus d'intérêt public, mais aussi la perte du droit de réclamer une révision pour des milliers d'utilisateurs en Amérique latine.

Au-delà du manque d'éléments pour déterminer pleinement chacune des raisons de ce changement de critères, le fait est qu'en 2020, les plateformes ont pris des décisions et modifié la manière et les processus dans lesquels elles modèrent les contenus. Ces modifications des processus et des normes communautaires qui les régissent ont marqué un changement radical par rapport à la manière dont Facebook, Twitter et YouTube traitaient auparavant le contenu créé par les utilisateurs.

Deux phénomènes semblent s'être produits cette année. Le premier, une augmentation très significative des messages dont le contenu est généralement considéré comme un «discours de haine » sur les réseaux sociaux, depuis la pandémie de COVID-19. Facebook est le réseau social sur lequel - du moins selon les données fournies par les plateformes elles-mêmes - la croissance a été la plus forte. Entre 2019 et 2020, les messages sur lesquels ce réseau social a agi, considérés comme des discours de haine, ont augmenté de près de 300 %. En analysant 2020 plus en détail, il est frappant de constater que cette croissance est beaucoup plus importante au cours du deuxième trimestre de l'année. Comme le souligne le chapitre précédent, à partir du mois de mars - date de l'explosion de la pandémie mondiale de COVID-19 - le nombre de posts modérés par la plateforme pour être considérés comme des discours de haine a doublé et est resté à ce niveau pendant le reste de l'année. Twitter et YouTube ont également connu une croissance, mais pas aussi importante.

Le deuxième phénomène à souligner est le fait que, suite aux effets de cette augmentation des discours de haine et aux plaintes de la société civile à cet égard, Facebook, Twitter et YouTube ont décidé d'approfondir leur surveillance et leur intervention et d'élargir les types de contenus qu'ils considèrent comme hors de leurs normes communautaires. Cependant, il ne semble pas évident pour de nombreux analystes dans le monde que ces mesures soient suffisantes ou adéquates, et la manière dont elles ont été mises en œuvre pose des problèmes majeurs, qui affectent des droits fondamentaux.

Malgré une croyance populaire, encore largement répandue, les réseaux sociaux n'ont jamais été des espaces d'échange totalement ouverts ou « non réglementés ». Depuis des années, les plateformes modèrent les contenus qu'elles considèrent comme «illégaux », mais aussi ceux qui répondent à des caractérisations encore plus vagues (et non interdites par la loi), comme ce qu'elles considèrent comme indécent, obscène et contraire aux bonnes mœurs de leur pays d'origine.

L'arrivée de la COVID-19, une pandémie mondiale qui a conduit des millions de personnes à s'isoler chez elles, à réduire leurs contacts et à travailler à distance, a eu toutes sortes d'impacts. L'un d'entre eux était l'augmentation des discours haineux sur les plateformes sociales, mais un autre, peut-être moins détectable à première vue, était le changement dans les processus de modération qui sont effectués sur le contenu que les utilisateurs publient. Selon une recherche menée sur la plateforme de recherche Crowdtangle (qui suit l'utilisation de hashtags ou de mots sur Facebook, Instagram et Twitter), entre février 2020 et mars 2021, 43 779 posts ont été générés sur Facebook utilisant l'expression « virus chinois » et ont enregistré un total de 3 535 409 interactions. Les deux moments où cette expression a été le plus utilisée ont été en mars et avril 2020.

Les gouvernements du monde entier ont demandé une distanciation sociale soutenue de leurs citoyens et les plateformes ont été contraintes d'envoyer des milliers de modérateurs humains chez eux. Cette décision a entraîné une augmentation significative de l'utilisation d'outils automatisés et de l'intelligence artificielle dans l'examen des millions de messages qui sont téléchargés sur les médias sociaux chaque minute. Bien qu'en constante amélioration, ces systèmes automatisés ne sont pas encore en mesure de comprendre les différences de langage, de langue, d'idiosyncrasie et de culture de millions d'utilisateurs dans le monde, ainsi que l'importance du contexte dans la définition de concepts aussi complexes que les discours de haine.

Selon l'étude de l'Unesco intitulée Countering Online Hate Speech (Combattre les discours de haine sur Internet), il existe au moins cinq approches non législatives possibles du problème des discours de haine en ligne, qui font toutes directement allusion au rôle des plateformes en tant que partie substantielle de la solution au problème. Dans ce document, l'Unesco propose un suivi et une analyse par la société civile, la promotion par les internautes du contre-discours en «peer-to-peer», une série de mesures prises par des ONG afin d'informer les autorités sur certains cas, la création de campagnes pour favoriser des actions menées par les fournisseurs Internet qui hébergent les contenus spécifiques, et la responsabilisation des utilisateurs à travers l'éducation et la formation concernant les connaissances, les compétences et les aspects éthiques de l'exercice de la liberté d'expression sur Internet.

Il est également clair que des erreurs dans la détection des discours de haine sur les plateformes peuvent entraîner la suppression de contenus qui ne relèvent pas de cette définition et donc un impact important sur la liberté d'expression en tant que droit humain fondamental.

Les plateformes se sont développées de manière exponentielle dans le monde entier et sont devenues des espaces d'échange d'idées. Ce qui s'y passe a donc une incidence directe (ou potentielle) sur le traitement du débat public. Permettre aux gouvernements et aux plateformes de devenir des régulateurs de contenu peut avoir pour conséquence de réduire au silence les voix dissidentes, en particulier dans les sociétés autoritaires.

Mais, comme l'affirme Díaz Hernández, le problème n'est pas seulement que les interdictions entraînent des restrictions excessives ou disproportionnées de la liberté d'expression, mais aussi qu'elles sont souvent inefficaces pour traiter et résoudre le problème à la racine, car elles ne jouent pas le rôle de contrer les discours de haine mais exacerbent souvent le climat de violence et de polarisation sociale qui a donné naissance au contenu à sa source.

Il est également important de garder à l'esprit que les problèmes découlant de la réglementation des contenus sur les plateformes ne concernent pas seulement la réglementation des contenus eux-mêmes, mais aussi l'architecture de l'internet telle que nous la connaissons, ainsi que ses caractéristiques d'espaces théoriquement extra-spatiaux et extraterritoriaux. Sur la base de cette structure et du rôle que les plateformes et les réseaux sociaux jouent dans cet écosystème, chacun de ces environnements a ses propres règles de fonctionnement et a généré ses propres définitions de ce qui est ou n'est pas interdit et autorisé. En ce sens, une partie du problème réside dans le fait qu'il ne s'agit pas seulement de savoir ce que la législation de chaque État entend par discours de haine, mais ce que ce terme signifie pour Facebook, Twitter ou YouTube, lorsque ceux-ci ne sont pas soumis à des contrôles démocratiques et n'offrent pas de garanties de procédure régulière ou de transparence, entre autres.

La pandémie mondiale a eu toutes sortes d'impacts sur la vie des gens. L'un de ces impacts est peut-être aussi le début d'une discussion sur le rôle des plateformes en tant que modératrices de contenu, sur les problèmes que pose le fait de les autoriser ou de les encourager à jouer le rôle de gardiennes de l'internet.

ANA LAURA PÉREZ

Uruguay

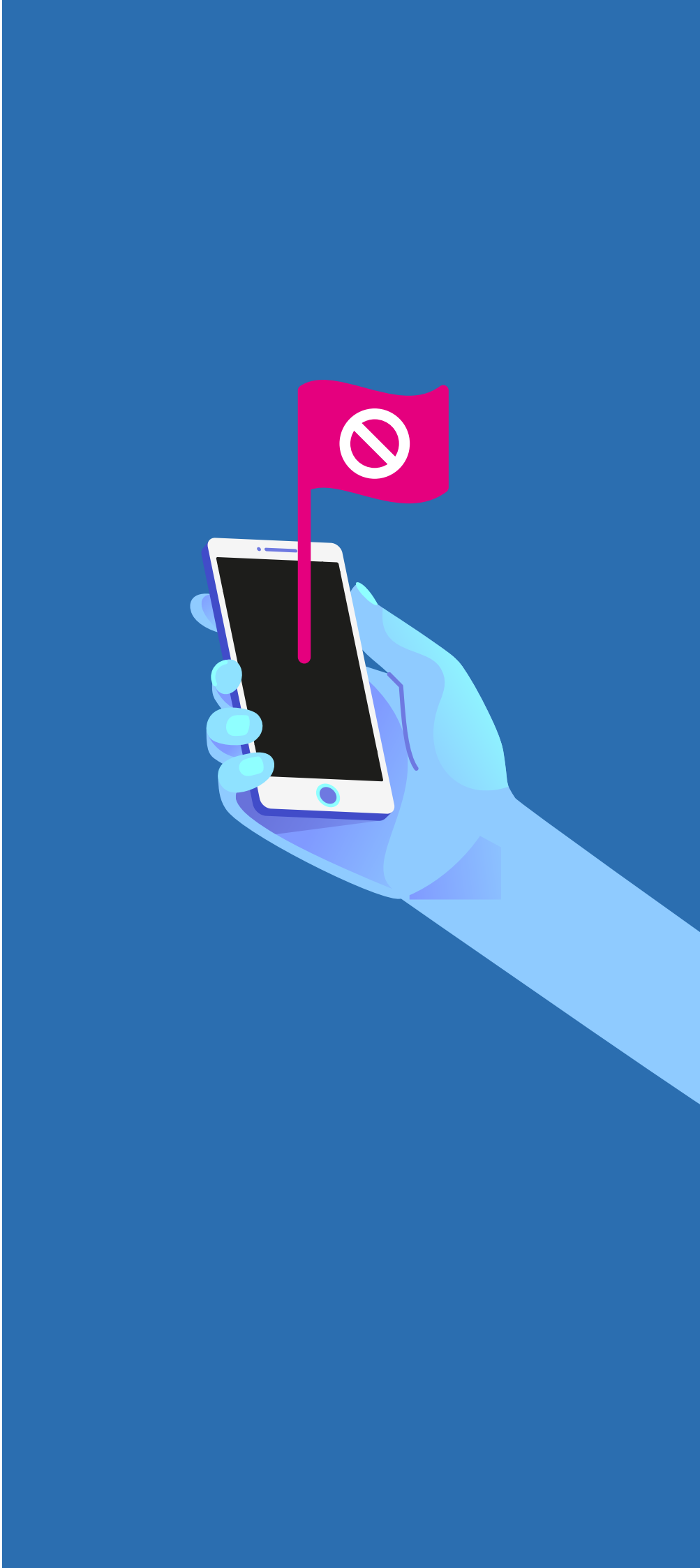
À PROPOS DE L'AUTEURE

Elle est titulaire d'un diplôme en communication et journalisme de l'université ORT, d'un diplôme en études latino-américaines de l'université de Montevideo et d'un master en administration des affaires de l'institut d'études commerciales de Montevideo.

Elle a travaillé pendant 20 ans comme journaliste et éditrice dans certains des médias les plus influents de son pays: les journaux El Observador et El País et l'hebdomadaire Búsqueda. Elle a également travaillé comme présentatrice et participante à des programmes télévisés sur TV Ciudad, la chaîne publique de la municipalité de Montevideo. Actuellement, elle est aussi responsable de Produit numérique du journal El País.

Elle a été coordinatrice de Journalisme et contenus numériques à la licence en communication de l'université ORT, où elle enseigne depuis près de dix ans.

Elle a participé en tant que conférencière, oratrice et panéliste à divers événements sur le journalisme, notamment sur la désinformation et les plateformes numériques, sujets sur lesquels elle s'est spécialisée ces dernières années et sur lesquels elle a donné des cours de formation à des journalistes en Uruguay et dans plusieurs pays d'Amérique latine.



Financé par l'Union Européenne