

Lutter contre les discours de haine sur les réseaux sociaux : difficultés contemporaines

Document de synthèse

01

Le présent document a été produit par des chercheurs de l'Oxford Internet Institute, avec le soutien de l'UNESCO, en tant que contribution à la Stratégie et au Plan d'action des Nations Unies pour la lutte contre les discours de haine, dans le cadre du projet « #CoronavirusFacts : Lutte contre la « désinfodémie » sur la COVID 19 dans les environnements sujets aux conflits », financé par l'Union européenne. Il s'inscrit dans le cadre de la collaboration de l'Oxford Internet Institute avec l'UNESCO, qui vise à élaborer une boîte à outils qui cartographie les méthodes, ressources et projets de recherche existants développés pour surveiller la présence, la propagation et l'impact des discours de haine en ligne, ainsi que pour évaluer les capacités et les pratiques destinées à les combattre. Tous les commentaires relatifs à ce document de discussion qui sont susceptibles de contribuer à éclairer cette étude plus large sont les bienvenus.

La lutte contre les discours de haine est un effort déployé sur plusieurs niveaux, qui consiste à s'attaquer à leurs causes profondes et à leurs moteurs, à les empêcher de se traduire en violences et à faire face à leurs conséquences sociétales plus vastes. Pour proposer des réponses efficaces aux discours de haine, y compris par le biais de l'éducation, il est essentiel de mieux surveiller et analyser le phénomène en s'appuyant sur des données claires et fiables. À l'ère du numérique, cela suppose également de mieux comprendre l'apparition, la virulence et la portée des discours de haine en ligne.

L'identification des discours de haine en ligne à des fins de recherche se heurte à de nombreuses difficultés sur le plan méthodologique, telles que le choix des définitions utilisées pour délimiter le problème, les contextes sociaux et historiques, les subtilités linguistiques, la variété des communautés en ligne et les formes de discours de haine en ligne (type de langage, images, etc.). D'un point de vue technologique, les discours de haine en ligne sont difficiles à étudier en raison de la fiabilité variable des systèmes de détection, de la nature opaque des algorithmes propriétaires et du manque d'accès aux données détenues par les entreprises, entre autres. Il est indispensable de clarifier la manière dont ces difficultés peuvent être résolues afin de mieux comprendre comment les discours de haine en ligne apparaissent et prolifèrent, et de formuler par la suite des réponses efficaces.

La Stratégie et le Plan d'action des Nations Unies identifient une série de domaines prioritaires pour le suivi et l'analyse des discours de haine, stipulant que les entités compétentes des Nations Unies devraient être en mesure « de dégager, de surveiller et d'analyser les tendances en matière de discours de haine, et de collecter des données à ce sujet ». Lorsqu'elles se concentrent sur les discours de haine en ligne, les entités des Nations Unies sont encouragées à promouvoir « la recherche sur la relation entre l'utilisation abusive d'Internet et des médias sociaux pour diffuser des propos haineux et les facteurs qui poussent certaines personnes à la violence », à « cartographier les opportunités et les risques émergents liés à la propagation des discours de haine que posent les nouvelles technologies et les plates-formes numériques » et, enfin, à « définir des protocoles d'action qui tiennent compte des nouvelles formes de discours de haine numériques ».

Au cours de l'année écoulée, la pandémie de COVID-19 a mis en lumière la pertinence de la Stratégie et du Plan d'action des Nations Unies pour la lutte contre les discours de haine, dont une véritable vague a déferlé sur le monde entier, exacerbant encore un peu plus l'intolérance et la discrimination envers certains groupes et déstabilisant les sociétés et leurs systèmes politiques.

Le présent document de discussion vise à donner un aperçu des principaux aspects qu'il convient de prendre en considération pour lutter contre l'apparition de discours de haine sur les réseaux sociaux, que ce soit par le biais de l'adoption de règlements concrets par les entreprises concernées, d'efforts de contre-attaque, de législations ou de mesures éducatives préventives. Le document est divisé en trois parties : la première se concentre sur les définitions de « discours de haine » et les cadres juridiques connexes, la deuxième partie présente des outils et des techniques de surveillance des discours de haine en ligne et évoque les mesures de la prévalence des discours de haine en ligne, et la troisième aborde les mesures potentielles de prévention et de contre-attaque.

1

DÉFINIR LE DISCOURS DE HAINE

Les difficultés liées à la lutte contre les discours de haine et à l'établissement d'une législation en la matière commencent avec la définition même du phénomène. Il n'existe pas de définition internationalement acceptée de ce qu'est un discours de haine. Par ailleurs, le sujet couvre plusieurs domaines juridiques tels que la liberté d'opinion et d'expression, la discrimination et les appels ou incitations à la discrimination, à l'hostilité ou à la violence.

Le « ***Dangerous Speech Project*** » de Susan Benesch¹ suggère que le terme « discours de haine » pose deux grandes difficultés. Tout d'abord, « haine » est un terme vague qui peut renvoyer à différents niveaux d'intensité et entraîner des conséquences différentes : « la « haine » d'un discours de haine signifie-t-elle que l'orateur hait les autres, qu'il cherche à persuader les autres de haïr ou qu'il souhaite que les autres se sentent haïs ? »². Ensuite, « discours de haine » signifie fondamentalement qu'une personne ou un groupe est visé en raison de son identité ou de son appartenance à un groupe. Cela nécessite qu'une loi ou une définition précise si toutes les identités et tous les groupes relèvent de cette loi et, dans le cas contraire, quels types de groupes sont inclus. Le « Dangerous Speech Project » soutient que des lois trop larges peuvent être utilisées à tort contre des groupes vulnérables ou l'opposition politique et civique, ce qui peut parfois nuire aux groupes que les lois sur les discours de haine ont précisément vocation à protéger. Néanmoins, il est également possible de faire valoir qu'une définition trop étroitement axée sur des groupes et des identités spécifiques pourrait conduire à des exclusions juridiques ou à l'absence d'outils juridiques susceptibles de lutter contre le problème.

Bien que la portée du présent document de discussion ne permette pas d'étudier ces difficultés en détail, un examen des lois internationales et nationales à travers le monde illustre les complexités et les différentes interprétations qui peuvent être appliquées aux discours de haine.

Au niveau mondial, parallèlement à la **Déclaration universelle des droits de l'homme**, non contraignante, le **Pacte international relatif aux droits civils et politiques** (PIDCP) fait suivre la liberté d'expression (article 19) par l'interdiction de tout appel à la haine qui constitue une incitation à la discrimination, à l'hostilité ou à la violence (article 20). Les articles 19 et 20 posent également des limites aux restrictions de la liberté d'expression – celles-ci devant être « expressément **fixées par la loi** et [...] **nécessaires : (a) au respect des droits ou de la réputation d'autrui ; (b) à la sauvegarde de la sécurité nationale, de l'ordre public, de la santé ou de la moralité publiques** ».

En complément de ces principes, le Plan d'action de Rabat propose un « **examen de seuil comportant six étapes** » afin de justifier les restrictions à la liberté d'expression, compte tenu du **contexte sociopolitique, du statut de l'orateur, de l'intention d'inciter à l'antagonisme, du contenu du discours, de la portée de sa diffusion et de la probabilité de préjudice**.

Autre texte important dans le domaine des discours de haine, la *Convention internationale sur l'élimination de toutes les formes de discrimination raciale* s'avère plus stricte que l'article 20 du PIDCP en n'exigeant pas d'intention ni d'« appel à la haine » et en incluant la diffusion dans sa liste des pratiques punissables par la loi. La *Convention pour la prévention et la répression du crime de génocide* et la *Convention sur l'élimination de toutes les formes de discrimination à l'égard des femmes*, notamment, sont également pertinentes en la matière.

L'organisation pour la liberté d'expression, ARTICLE 19, a élaboré les *Principes de Camden sur la liberté d'expression et l'égalité* sur la base de discussions avec des fonctionnaires de l'ONU et des experts du monde universitaire et de la société civile. Ces Principes fournissent des orientations interprétatives sur les articles du PIDCP et visent à dissuader les acteurs d'abuser de l'article 20 en apportant des précisions sur l'« incitation » et ce qui constitue des faits de « discrimination », d'« hostilité » et de « violence ».

¹ The Dangerous Speech Project. 2021. <https://dangerousspeech.org/>

² Benesch, S. 2021. *Dangerous Speech: A Practical Guide*.

The Dangerous Speech Project, 2021, p. 7. <https://dangerousspeech.org/>

Lorsqu'il transpose le droit et les principes internationaux dans la législation nationale, chaque pays adopte une définition légèrement différente des discours de haine et de leur mode d'expression, de leurs cibles potentielles et du type de préjudice qui doit être constaté pour que les discours eux-mêmes soient considérés comme haineux. L'absence de définition unifiée est l'une des principales difficultés à résoudre pour lutter contre les discours de haine en ligne, qui ne s'arrêtent pas nécessairement aux frontières des pays.

Définir les discours de haine est également important pour la recherche et les efforts de plaidoyer, en particulier lorsqu'il s'agit de déterminer leurs conséquences sociétales. Les préjudices résultant des discours de haine peuvent toucher les individus (préjudices psychologiques), les groupes et les communautés, et la société (érosion des droits et des biens publics). Étant donné que les discours de haine ciblent les personnes en raison des caractéristiques des groupes auxquels elles appartiennent, l'analyse du préjudice au niveau communautaire revêt une importance particulière. Les préjudices dus aux discours de haine sont *répartis inégalement* parmi la population en général, les groupes marginalisés étant les plus lourdement touchés. Ils sont également *cumulatifs* pour ceux qui les subissent, l'expérience antérieure des discours de haine étant une variable clé dans l'estimation du préjudice découlant du fait d'être la cible d'un discours de haine.³

DISCOURS DE HAINE EN LIGNE

Les discours de haine en ligne ne sont pas intrinsèquement différents des discours de haine hors ligne. Ils diffèrent toutefois par la nature des interactions dans le cadre desquelles ils s'inscrivent et par l'utilisation et la diffusion de mots spécifiques, d'accusations et de théories complotistes qui peuvent évoluer, atteindre des sommets et disparaître très rapidement. Un message de haine peut devenir viral en quelques heures, voire quelques minutes.

Le rapport de l'UNESCO intitulé *Combattre les discours de haine sur Internet*, paru en 2015, souligne la facilité avec laquelle les discours de haine en ligne peuvent être produits et diffusés à moindre coût, sans passer par un

processus d'édition comme d'autres formes de travail écrit, connaissent des niveaux d'exposition variables suivant la popularité des publications, et peuvent être publiés par-delà les frontières dans la mesure où les serveurs des plates-formes et les sièges des réseaux sociaux ne sont pas nécessairement dans le même pays que les utilisateurs et le public visé. Les discours de haine en ligne peuvent également être disponibles plus longtemps et connaître des vagues de popularité, être associés à de nouveaux réseaux ou réapparaître, mais aussi rester anonymes. Les questions liées à la responsabilité de la modération des espaces et aux critères permettant de déterminer si et quand un contenu doit être supprimé ont par conséquent fait l'objet de vastes débats.

Ces débats ont été illustrés par l'adoption de lois telles que la **NetzDG (pour l'application du droit sur les réseaux sociaux)** en Allemagne. Présentée en 2017, celle-ci exige que les plates-formes des réseaux sociaux regroupant plus de 2 millions d'utilisateurs mettent en œuvre une procédure transparente de modération des contenus illégaux (dont les discours de haine), retirent les contenus considérés comme étant illégaux dans les 24 heures et rendent compte régulièrement des mesures prises. Cette loi a fait l'objet de vives critiques l'accusant d'obliger les plates-formes à jouer le rôle de « censeurs privés » pour des décisions qui devraient être prises par les tribunaux, et de les pousser, compte tenu de la brièveté des délais et du montant élevé des amendes, à retirer des quantités excessives de contenus afin d'éviter de s'exposer à de lourdes sanctions. En 2020, la loi a été révisée afin d'exiger des plates-formes de réseaux sociaux qu'elles transmettent les contenus jugés illégaux au Bureau fédéral de la police criminelle. Une autre révision presque simultanée a également renforcé les droits des utilisateurs en exigeant des plates-formes qu'elles facilitent le signalement des contenus illégaux et permettent de faire appel de la décision de suppression ou de maintien d'une publication.

Dans le même ordre d'idées, l'élaboration de lois pour lutter contre les discours de haine en ligne et hors ligne est souvent alourdie par des processus d'examen compliqués dus aux difficultés de définition et à la nécessité de respecter la liberté d'expression dans le cadre de la loi. Il est par conséquent nécessaire de déployer des méthodes qui dépassent le cadre des mesures juridiques pour lutter contre les discours de haine.

³ Gelber, K. and McNamara, L. 2016. Evidencing the harms of hate speech. *Social Identities*, Vol. 22, No. 3, pp. 324-341.

2

OUTILS ET TECHNIQUES DE MESURE ET DE SURVEILLANCE DES DISCOURS DE HAINE

Les politiques et les outils permettant la détection, la surveillance et la modération des discours de haine en ligne varient selon les contextes, les acteurs et les plateformes.

Les méthodes de détection peuvent être regroupées en deux catégories : les efforts plus complets qui s'appuyaient initialement sur des filtres de mots-clés et des méthodes participatives ; et ceux qui font appel à des modérateurs, qui examinent les contenus signalés par les utilisateurs comme étant des discours de haine et qui décident s'ils doivent être considérés comme tels ou non. Bien qu'une approche manuelle présente l'avantage de saisir le contexte et de réagir rapidement aux changements, le processus est chronophage et coûteux et demande beaucoup de main-d'œuvre, ce qui limite son évolutivité et la rapidité des solutions. En raison de ces difficultés, du volume croissant de contenus produits sur les réseaux sociaux et des progrès de l'apprentissage automatique et des techniques de traitement du langage naturel, les plateformes et les chercheurs ont développé des solutions de détection automatisées et les utilisent de plus en plus. De nombreuses initiatives plus récentes font appel à différentes méthodes combinées. Nous proposons ici la définition de quelques termes clés liés à ces méthodes :

- **Apprentissage automatique** : techniques qui utilisent des algorithmes informatiques pouvant s'améliorer automatiquement grâce à l'expérience et à l'utilisation des données.
- **Traitement du langage naturel** : techniques qui traitent et analysent de grandes quantités de données en langage naturel.
- **Approches basées sur des mots-clés** : méthodes utilisant une ontologie ou un dictionnaire et identifiant des textes susceptibles de contenir des mots clés potentiellement haineux.
- **Sémantique distributionnelle** : méthodes de quantification et de catégorisation des similitudes entre les mots, les expressions et les phrases en fonction de la façon dont elles sont distribuées dans de grands échantillons de données.
- **Analyse de sentiments** : méthodes visant à expliquer quels types d'attitudes sont transmis par rapport à un sujet dans un texte donné.
- **Métadonnées source** : certaines méthodes alimentent les modèles via les méta informations des données, telles que les données sur les utilisateurs associés aux messages, y compris les fonctionnalités basées sur le réseau telles que le nombre d'abonnés.
- **Apprentissage profond** : classe d'algorithmes d'apprentissage automatique qui utilise plusieurs couches pour extraire progressivement des fonctionnalités de niveau supérieur des entrées brutes.

Les entreprises des réseaux sociaux sont majoritairement passées de la réaction face aux publications signalées par les utilisateurs comme étant des discours de haine à une détection et une gestion proactives de ces contenus par le biais de leurs systèmes automatisés, avant même que les utilisateurs aient pu les voir. Si elles restent nécessaires pour lutter contre les discours de haine à grande échelle, ces méthodes entraînent également des complications : la détection automatisée des discours de haine fait des erreurs et il est possible que des contenus non haineux soient supprimés. Le retrait excessif de contenus risque d'avoir un effet dissuasif et de nuire à la liberté d'expression.

Pour améliorer les activités de surveillance, de nouveaux outils de détection des discours de haine sont continuellement mis au point. **L'interface de programmation Perspective⁴**, par exemple, est un outil open source de Jigsaw (un incubateur de Google) et de l'équipe de Google chargée des technologies de lutte contre les abus, utilisé par les organismes de presse et les produits de Google. Disponible en sept langues (allemand, anglais, français, espagnol, italien, portugais et russe), cet outil utilise l'apprentissage automatique pour noter des expressions en fonction de leur toxicité potentielle dans une conversation. **Facebook** a déclaré que la dernière version de son outil de surveillance et de détection des discours de haine sur ses plates formes avait amélioré la compréhension sémantique du langage et la compréhension des contenus, dont l'analyse des images et des commentaires, entre autres éléments.⁵

Des chercheurs et des organisations de la société civile ont également travaillé à la mise au point d'outils de détection des discours de haine. En voici quelques exemples :

- La plate-forme kényane **Umati** a été l'une des premières à surveiller manuellement les publications en ligne rédigées dans les langues pertinentes au Kenya.
- Davidson et al. (2017) ont développé l'outil **Hate Sonar** en se basant sur un modèle de régression logistique formé grâce aux données des forums sur Internet et de Twitter.
- ADL et D-Lab, de l'Université de Californie à Berkeley, ont élaboré **Online Hate Index**, un indice de haine en ligne conçu pour transformer la compréhension humaine des discours de haine grâce à l'apprentissage

automatique en un outil évolutif qui peut être déployé sur les contenus d'Internet afin de découvrir la portée et l'étendue des discours de haine en ligne.

- L'équipe du projet « Measures & Counter-measures » de l'Alan Turing Institute a développé un outil utilisant des méthodes d'apprentissage profond **pour détecter les préjugés sur l'Asie de l'Est** sur les réseaux sociaux.
- Moon et al. (2020) ont développé un outil de **détection des discours de haine en coréen**, dont le modèle repose sur l'ajout d'une qualification « parti pris » en plus de « haine ».
- **Hatemeter** détecte les discours de haine antimusulmans à l'aide de techniques d'apprentissage automatique et de traitement du langage naturel. Cette plate-forme est disponible en anglais, français et italien.
- **COSMOS** recueille et analyse les données de Twitter en temps réel par spécification de mots clés, en utilisant l'analyse de sentiments et le traitement du langage naturel.
- **MANDOLA** détecte les contenus haineux grâce à une combinaison d'analyse de sentiments, de traitement du langage naturel, d'apprentissage automatique et d'apprentissage profond.

Il est important de noter que la surveillance des discours de haine en ligne dépend de l'accès aux données, surtout pour ce qui concerne les plates-formes des réseaux sociaux. Par ailleurs, nombre des outils existants sont aujourd'hui monolingues et souvent limités à la langue anglaise, et de plus amples recherches sur les performances des méthodes de détection multilingues sont nécessaires. En outre, la grande majorité des recherches et de la surveillance des discours de haine sur les plates-formes des réseaux sociaux s'est focalisée sur les États Unis et l'Europe. Cela a entraîné des écarts, non seulement au niveau des outils et des données, mais aussi de la compréhension de l'ampleur et de la dynamique de la propagation des discours de haine dans les autres régions du monde. Étant donné la nature contextuelle des discours de haine, il est crucial de combler ces écarts.

⁴ Pour de plus amples informations sur Perspective, rendez-vous sur <https://www.perspectiveapi.com/>

⁵ Source : <https://ai.facebook.com/blog/ai-advances-to-better-detect-hate-speech/>

PRÉVALENCE DES DISCOURS DE HAINE SUR LES PLATES-FORMES DES RÉSEAUX SOCIAUX

Grâce à des outils de détection automatisés basés sur les méthodes disponibles aujourd'hui, Twitter, Facebook, Instagram et YouTube font état d'un accroissement de la quantité de contenus signalés et/ou supprimés. Entre janvier et mars 2021, YouTube a retiré 85 247 vidéos qui enfreignaient sa politique relative aux discours de haine. Ses deux rapports précédents montraient des chiffres similaires. Pour le même trimestre, Facebook et Instagram ont respectivement signalé un total de 25,2 millions et 6,3 millions de contenus ayant déclenché une action. Selon le dernier rapport de Twitter au sujet de la transparence, l'entreprise a retiré 1 628 281 contenus considérés comme contraires à sa politique relative aux discours de haine entre juillet et décembre 2020.

Sur les plates-formes des réseaux sociaux, la prévalence des discours de haine est déterminée sur la base d'un échantillon des contenus vus par les utilisateurs. En d'autres termes, ce qui est décompté est une part approximative de ce qu'il reste des discours de haine sur la plate forme, une fois le travail de détection et de suppression proactives effectué par l'entreprise. À ce jour, Facebook est la seule plate-forme qui fournit des données sur la prévalence. L'entreprise a signalé qu'entre janvier et mars 2021, la prévalence des discours de haine était comprise entre 0,05 % et 0,06 %, soit une légère diminution par rapport à ses deux rapports précédents. Certaines études indiquent que la prévalence des discours de haine sur les plates-formes traditionnelles telles que Twitter et Wikipédia représente moins de 1 % de l'ensemble des contenus, tandis que sur les plates formes alternatives plus spécialisées, telles que Gab et 4chan, entre 5 % et 8 % des contenus hébergés peuvent être injurieux⁶. Les preuves de la prévalence des discours de haine sur les plates-formes des réseaux sociaux restent incomplètes, ce qui est en partie attribuable au manque de transparence et à l'absence d'accès aux données des plates-formes.

⁶ Zannettou et al. 2018; Mathew et al. 2018; Hine et al. 2017; Vidgen et al. 2019.

3

COMBATTRE LES DISCOURS DE HAINE SUR INTERNET

Soulignons tout d'abord que la lutte contre les discours de haine (et, par extension, la sélection des outils, des stratégies et des efforts de prévention appropriés) est compliquée par plusieurs facteurs. Les réponses des différents acteurs aux questions clés dans un éventail de contextes s'avèrent peu consensuelles. De quelle manière les discours de haine nuisent-ils et à quel moment ce préjudice est-il suffisamment grave pour justifier leur régulation ? De façon plus granulaire, quels types de préjudices, liés à quels comportements renvoyant aux discours de haine, justifient une régulation conforme aux normes et lois internationales relatives aux droits de l'homme et à la liberté d'expression ?

L'architecture d'Internet ajoute d'autres obstacles distincts à la lutte contre les discours de haine, tels que la permanence, l'itinérance, l'anonymat et le caractère interjuridictionnel des contenus en ligne, l'éventail d'architectures des plates-formes et la nature hétérogène du système de gouvernance d'Internet, qui repose sur de multiples intervenants.

Malgré ces difficultés, de nombreux groupes et individus s'engagent de diverses façons pour lutter contre les discours de haine en ligne et renforcer de manière préventive la résilience des internautes face à ces phénomènes.

RECOURS JURIDIQUES DES ÉTATS

Les recours juridiques représentent un moyen fondamental de combattre les discours de haine. Bien que les positions sur les discours de haine varient d'une région à l'autre et continuent d'évoluer au fur et à mesure que la compréhension du sujet s'améliore, il existe un certain nombre de principes internationaux, d'accords régionaux, de lois nationales et d'exemples

de jurisprudence conformes aux normes internationales relatives aux droits de l'homme qui contiennent des clauses sur les discours de haine en ligne et hors ligne, comme indiqué au début du présent document.

Cependant, les réponses strictement juridiques aux discours de haine en ligne font rapidement apparaître certains problèmes, tels que les préoccupations relatives à l'équilibre des droits, la possibilité que des intervenants puissants abusent des restrictions des droits, la définition d'un seuil fiable pour identifier l'incitation à la violence, et la mauvaise compréhension du rapport entre discours de haine et violence hors ligne. Plus important encore dans la lutte contre les discours de haine en ligne, le principal problème des recours juridiques est la limitation de l'autorité individuelle de chaque État sur les espaces numériques en ligne. Pour être efficace, la lutte contre les discours de haine en ligne ne peut exclusivement reposer sur des recours juridiques nationaux⁷.

En 2016, plusieurs grandes entreprises du secteur des technologies sont convenues d'adopter le Code de conduite de la Commission européenne visant à combattre les discours de haine en ligne, qui exige que ces entreprises examinent les discours de haine dans les 24 heures suivant la réception d'un signalement. Si cette approche n'est pas sans poser des difficultés en raison d'une grande divergence en termes de services et de définitions opérationnelles des discours de haine, elle constitue néanmoins un effort important pour encourager la collaboration et faire le lien entre les approches juridiques et extrajuridiques dans le domaine des discours de haine.

⁷ Depuis 2013, l'initiative de l'UNESCO pour la formation des juges a contribué à renforcer les capacités des acteurs judiciaires en matière de normes internationales et régionales sur la liberté d'expression, l'accès à l'information et la sécurité des journalistes dans les régions du monde entier, en particulier motivée par le fait qu'une meilleure façon de traiter les cas de discours de haine est l'un des intérêts majeurs de nombreux opérateurs judiciaires. Plus de 23 000 acteurs du secteur judiciaire ont été formés sur ces questions, notamment grâce à une série de cours en ligne ouverts à tous (MOOC), à des ateliers et des formations sur le terrain et à la publication de plusieurs boîtes à outils et directives.

RÉPONSES DES ENTREPRISES DU SECTEUR DES TECHNOLOGIES

En 2021, YouTube et Facebook ont fait état d'une augmentation des contenus identifiés et signalés, ainsi que d'une plus grande proportion de contenus signalés par l'entreprise elle-même par rapport aux contenus signalés par les utilisateurs. Ce constat s'explique par l'utilisation croissante de systèmes de détection automatisés. Cette tendance s'accompagne toutefois d'une augmentation des contenus restaurés par rapport aux périodes précédentes. Entre janvier et mars 2021, Facebook et Instagram ont restauré respectivement 408 700 et 43 700 contenus. Bien que les rapports suggèrent que les plates-formes prennent de plus en plus de mesures contre les contenus haineux, nous ignorons si cela est dû à une augmentation du nombre d'abus, au renforcement de la rigueur des politiques des plates-formes ou à une augmentation du nombre de faux positifs.

Les entreprises de réseaux sociaux étant basées dans des juridictions nationales, elles sont directement influencées par le droit national et s'avèrent généralement plus sensibles aux demandes de celui-ci en matière de lutte contre les discours de haine. N'étant pour leur part pas liées par la territorialité, les plates-formes des réseaux sociaux ne peuvent que respecter leurs propres conditions de service, qui peuvent être plus strictes ou plus souples que les normes établies par les accords internationaux décrits plus haut. Les mesures prises par les plates-formes des réseaux sociaux comprennent le retrait des contenus considérés comme des discours de haine, l'envoi d'avertissements aux utilisateurs qui publient des discours de haine, la restriction de leur activité sur la plate-forme ou leur bannissement. Ces normes communautaires sont en constante évolution, notamment en ce qui concerne le recours de plus en plus fréquent à des méthodes de modération automatisées aux dépens des méthodes « humaines ».

À la lumière de ces difficultés, un mouvement multipartite appelant à une plus grande transparence des entreprises d'Internet afin d'améliorer leur responsabilité a pris de l'ampleur ces dernières années.

Il a notamment été proposé de prendre des mesures juridiques et réglementaires dans 30 pays et régions au moins, notamment par le biais de la loi européenne sur les services numériques, en cours d'élaboration. Les entreprises ont également pris des mesures afin de gagner en transparence. En 2021, Access Now a indexé plus de 70 entreprises qui émettent régulièrement des rapports sur la transparence⁸, même si ce nombre devrait être bien supérieur.

Le dossier d'information de l'UNESCO intitulé « Letting the Sun Shine In: Transparency and Accountability in the Digital Age » (Laissons entrer le soleil : transparence et responsabilisation à l'ère du numérique) présente le renforcement de la transparence comme une troisième voie entre l'excès de réglementation des contenus par les États, qui a entraîné des restrictions disproportionnées des droits de l'homme, et une approche de laissez-faire qui n'a pas permis de traiter efficacement les contenus problématiques tels que les discours de haine et la désinformation. Ce dossier propose 26 principes de haut niveau couvrant des questions qui concernent les contenus et les processus, la diligence raisonnable et la réparation, l'autonomisation, les dimensions commerciales, la collecte et l'utilisation des données à caractère personnel, et l'accès aux données.

RÉPONSES EXTRAJURIDIQUES ET INTERVENTIONS PRÉVENTIVES

D'autres réponses extrajuridiques proviennent de la recherche et des efforts de plaidoyer de la société civile ou peuvent se concentrer sur des mesures préventives qui renforcent la résilience des utilisateurs en ligne face aux discours de haine. Certaines de ces initiatives ciblent directement les causes et les conséquences des discours de haine en ligne, y compris par le biais de l'éducation, tandis que d'autres appellent à l'amélioration des mesures juridiques et technologiques.

Souvent axées sur la prévention à long terme, les initiatives fondées sur l'éducation sont au cœur de ces efforts. Les interventions éducatives peuvent servir à sensibiliser aux conséquences néfastes des discours de haine, à s'attaquer à leurs causes profondes et à alerter

⁸ <https://www.accessnow.org/transparency-reporting-index/>

efficacement sur les techniques de manipulation et de rhétorique utilisées pour répandre la haine, en ligne et hors ligne. En particulier, des programmes d'éducation aux médias et à l'information ont été développés et mis en œuvre dans le monde entier dans le but de donner aux internautes les compétences nécessaires pour examiner de manière critique les informations des contenus en ligne et identifier la désinformation et les contenus inquiétants et haineux. De la même manière, des efforts de « contre-discours » visant à faire face aux discours de haine par le biais d'actions positives, telles que l'Initiative pour le courage civique en ligne menée par Facebook en Allemagne, au Royaume-Uni et en France en 2017, ont également été déployés. D'autres initiatives de la société civile mettent l'accent sur la promotion du changement par les plates formes. En juillet 2020, la campagne Stop Hate for Profit a réuni une coalition de plus de 1 200 entreprises du monde entier et appelé à un boycott publicitaire contre les principales plates-formes, à la modération des discours de haine et à une pause publicitaire sur les comptes encourageant la discrimination contre certains groupes. Cette campagne, venue s'ajouter au nombre croissant de voix appelant à lutter contre les discours de haine en ligne, en particulier à la lumière de l'intensification des discours prenant pour cible des groupes marginalisés pendant la pandémie de COVID-19, a incité plusieurs entreprises de réseaux sociaux à modifier leurs directives communautaires.

RECOMMANDATIONS

Dans le but d'éclairer une politique fondée sur des preuves afin de limiter les discours de haine en ligne et d'empêcher qu'ils ne se traduisent en violences, tout en protégeant la liberté d'expression, il est essentiel de reconnaître et de surveiller ce phénomène, de recueillir des données et d'analyser les tendances en matière de discours de haine afin d'élaborer les stratégies appropriées pour y remédier. Les recommandations ci-dessous visent à identifier les actions clés qui permettraient de relever les nouveaux défis dus à l'émergence de discours de haine viraux et, en particulier, de traiter leurs conséquences hors ligne sur la paix, la stabilité et la jouissance des droits de l'homme pour tous.

1. Promouvoir une définition inclusive des discours de haine qui respecte la liberté d'expression

- Veiller à ce que les définitions soient conformes aux normes internationales, notamment aux dispositions du PIDCP et du Plan d'action de Rabat.

2. Mettre en place des coalitions multipartites

- Encourager le partage de données et d'expertise entre les organisations de défense des droits de l'homme, les intermédiaires d'Internet et le public.

- Habilitier les parties prenantes, notamment les communautés locales, à surveiller et à détecter les discours de haine sur les réseaux sociaux, en fonction de leur contexte et de leur langue.
- Organiser des échanges multipartites sur les tendances en matière de discours de haine, leur apparition et la façon des combattre.
- Recommander aux plates-formes de développer, en collaboration avec des groupes d'experts et le public, des définitions et des procédures opérationnelles qui devraient dépasser l'Amérique du Nord et l'Europe occidentale pour inclure plus de pays dans le monde.

3. Recueillir des données et encourager les pratiques reposant sur les données ouvertes là où celles-ci sont déjà collectées, tout en respectant la protection des données à caractère personnel

- Recueillir des données qualitatives auprès des personnes ciblées par des discours de haine afin de mieux comprendre la portée et la nature des préjudices.
- Recommander aux opérateurs d'Internet d'améliorer leurs pratiques en matière de transparence, y compris en publiant ouvertement des données sur les plaintes contre les discours de haine et leur résolution, ainsi que sur la précision et le fonctionnement de leurs systèmes de modération de contenus.
- Soutenir le développement de méthodes et d'outils abordables, accessibles et conviviaux pouvant être utilisés pour surveiller et détecter les discours de haine dans des contextes multiculturels et multilingues, dans des délais qui permettent le déploiement de contre-mesures.

4. Encourager les plates-formes à proposer des recours efficaces aux personnes dont les contenus ont été supprimés

- Faciliter la collaboration entre les entreprises de réseaux sociaux et les groupes de la société civile œuvrant pour la défense des droits numériques afin d'assurer la conformité des processus de modération et de suppression de contenus avec les besoins de la communauté.

5. Renforcer la sensibilisation aux discours de haine, l'éducation aux médias et à l'information et les compétences numériques par le biais de programmes éducatifs

- Consacrer des fonds et des ressources à l'élaboration de programmes éducatifs qui favorisent la résilience face aux discours de haine, qui soient éclairés par les tendances actuelles en la matière et qui répondent aux défis connexes. Cette approche nécessite une collaboration étroite entre les entreprises de réseaux sociaux, les instituts de recherche et les intervenants du secteur de l'éducation.
- Privilégier les approches éducatives préventives qui alertent sur les effets néfastes des discours de haine en ligne et qui favorisent l'éducation aux médias et à l'information, parallèlement aux efforts d'atténuation et de contre-attaque.
- Établir et soutenir des partenariats entre les établissements d'enseignement et les entreprises de réseaux sociaux afin d'accroître l'accès aux informations et aux ressources permettant de lutter contre les discours de haine sur les plates formes des réseaux sociaux, via des campagnes de diffusion ciblées ou la réorientation des utilisateurs vers des ressources externes.

6. Soutenir les organisations œuvrant dans le domaine des discours de haine en ligne

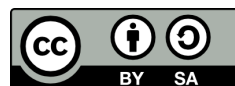
- Soutenir les organisations spécialisées qui se consacrent à la lutte contre les discours de haine et à leur surveillance, en particulier celles qui sont les mieux équipées pour prendre en compte les contextes locaux, et veiller à ce qu'elles bénéficient de ressources adéquates.

Ce document fait partie d'une série de documents de synthèse commandés et produits par l'UNESCO et le Bureau du Conseiller spécial pour la prévention du génocide des Nations Unies (OSAPG). Ces documents sont une contribution directe à la stratégie et le Plan d'action des Nations Unies et sont publiés dans le cadre du Forum multipartite et de la Conférence ministérielle sur la lutte contre les discours de haine par l'éducation, en septembre et octobre 2021.

L'apparition de la pandémie de COVID-19 a mis en évidence la pertinence de la Stratégie et le Plan d'action des Nations Unies, générant une vague de discours de haine dans le monde entier, exacerbant encore l'intolérance et la discrimination à l'égard de certains groupes et déstabilisant les sociétés et les systèmes politiques. Les documents de synthèse visent à analyser les questions clés liées à ce défi mondial et à proposer des réponses et des recommandations possibles.

Ce document a été commandé par la Section de la liberté d'expression et de la sécurité des journalistes de l'UNESCO dans le cadre du projet « #CoronavirusFacts, Addressing the 'Disinfodemic' on COVID-19 in conflict-prone environments » financé par l'Union européenne. Il a été rédigé par Jonathan Bright, Antonella Perini, Anne Ploin et Reja Wyss de l'Oxford Internet Institute de l'Université d'Oxford.

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
© UNESCO 2021



Œuvre publiée en libre accès sous la licence Attribution-ShareAlike 3.0 IGO (CC-BY-SA 3.0 IGO) (<http://creativecommons.org/licenses/by-sa/3.0/igo/>). Les utilisateurs du contenu du présent document acceptent les termes d'utilisation de l'Archive ouverte de libre accès UNESCO (www.unesco.org/open-access/terms-use-ccbysa-fr).

Titre original : *Addressing hate speech on social media: Contemporary challenges*

Publié en 2021 par l'Organisation des Nations Unies pour l'éducation, la science et la culture (UNESCO).

Les désignations employées dans ce document et la présentation des données qui y figurent n'impliquent de la part de l'UNESCO aucune prise de position quant au statut juridique des pays, territoires, villes ou zones, ou de leurs autorités, ni quant au tracé de leurs frontières ou limites.

Les idées et les opinions exprimées dans ce document sont celles des auteurs ; elles ne reflètent pas nécessairement les points de vue de l'UNESCO et n'engagent en aucune façon l'Organisation.

Ce document a été réalisé avec le soutien financier de l'Union européenne. Son contenu relève de la seule responsabilité des auteurs et ne reflète pas nécessairement les vues de l'Union européenne.

Création graphique : Dean Dorat

CI/FEJ/2021/DP/01