Platforms, Experts, Tools: Specialised Cyber-Activists Network

Rapport sur les exercices de monitoring

2019 – 2020

ar le programme « droits, égalité et
2014-2020) de l'Union Européenne

À propos du projet

Le projet **sCAN** – *Platforms, Experts, Tools: Specialised Cyber-Activists Network* (2018-2020), financé par l'UE et coordonné par la Licra (Ligue Internationale Contre le Racisme et l'Antisémitisme), a pour but de rassembler expertise, outils, méthodologie et connaissances concernant la haine en ligne et d'élaborer un ensemble de pratiques complet pour permettre d'identifier, d'analyser, de signaler et de réagir pour contrer les discours de haine en ligne. Ce projet s'appuie sur les résultats d'autres projets européens concluants, comme par exemple les projets « Research, Report, Remove: Countering Cyber-Hate phenomena » et « Facing Facts », et s'emploie à poursuivre, amplifier et renforcer les initiatives développées par la société civile en ce qui concerne la lutte contre les discours de haine.

Les partenaires du projet **sCAN** pourront, à travers une coopération européenne, renforcer et approfondir (davantage) leur fructueuse collaboration. Ils contribueront à la sélection et à l'apport d'outils de contrôle automatisés utiles pour un meilleur repérage du contenu haineux. Le projet s'attachera à renforcer les actions en termes de monitoring (comme les exercices de monitoring) instaurées par la Commission Européenne. Les partenaires rassembleront également leurs connaissances et observations respectives afin de mieux pouvoir identifier, expliquer et comprendre les tendances de la haine en ligne à l'échelle internationale. Le projet vise en outre à développer les moyens de l'Europe en proposant des cours en ligne pour les cybermilitants, les modérateurs et les formateurs, à travers la plateforme en ligne de Facing Facts.

sCAN sera mis en œuvre par dix partenaires européens : ZARA, Zivilcourage und Anti-Rassismus-Arbeit (Autriche), CEJI-A Jewish contribution to an inclusive Europe (Belgique), Human Rights House Zagreb (Croatie), Romea (République Tchèque), Respect Zone et Licra, Ligue Internationale Contre le Racisme et l'Antisémitisme (France), jugendschutz.net (Allemagne), CESIE (Italie), le Latvian Centre for Human Rights (Lettonie), et l'Université de Ljubljana, Faculty of Social Sciences (Slovénie).

Le projet **sCAN** est financé par la direction générale de la justice et des consommateurs de la Commission Européenne, dans le cadre du programme de l'Union Européenne « droits, égalité et citoyenneté ».

Clause de non-responsabilité

Ce rapport annuel est financé par le programme « droits, égalité et citoyenneté » (2014-2020) de l'Union européenne.
Le contenu de cette analyse représente uniquement le point de vue de ses auteurs et est la seule responsabilité du consortium du projet sCAN. La Commission européenne n'est pas responsable de l'usage qui pourrait être fait des informations qui y figurent.

Projet financé par le programme « droits, égalité et citoyenneté » (2014-2020) de l'Union européenne

Sommaire

Introduction

Au cours de la deuxième année du projet, les organisations partenaires de sCAN ont participé à deux exercices de monitoring, l'un avec la Commission européenne et l'autre avec le Réseau international contre la cyberhaine (INACH) et le projet Open Code for Hate-Free Communication (OpCode). L'objectif de ces exercices était d'évaluer l'adhésion des sociétés informatiques Facebook, Twitter, YouTube et Instagram au Code de conduite pour la Lutte contre les Discours de Haine en Ligne, élaboré en 2016 par la Commission européenne. Les partenaires de sCAN ont déjà participé à des exercices organisés par la Commission européenne et l'INACH.

Dans le Code de conduite, les sociétés informatiques s'engagent à « examiner la majorité des notifications valides pour la suppression des discours de haine illégaux en moins de 24 heures » et à supprimer ou restreindre l'accès aux contenus qui violent leurs lignes directrices communautaires et/ou leur droit national. Comme il est impossible d'évaluer le délai d'examen d'un rapport pour les organisations externes, les partenaires de sCAN ont enregistré le moment où l'entreprise notifiée a pris des mesures ou fourni un retour d'information sur les signalisations.

Entre le 4 novembre 2019 et le 13 décembre 2019, les partenaires de sCAN ont participé au cinquième exercice de monitoring organisé par la Commission européenne depuis 2016. Au cours de ce monitoring, les partenaires ont signalé 635 cas de discours haineux illégaux en ligne aux sociétés informatiques Facebook, Twitter, YouTube, Instagram, Dailymotion et Jeuxvidéo.

Le 20 janvier 2020 et le 28 février 2020, le projet sCAN a coopéré à l'organisation d'un monitoring inattendu avec INACH et le projet OpCode. Le calendrier de cette surveillance a été choisi de manière à tenir compte de la durée du projet sCAN jusqu'à la fin du mois d'avril 2020. Au cours de cette surveillance, les partenaires de sCAN ont signalé 484 cas de discours haineux illégaux en ligne aux sociétés informatiques Facebook, Twitter, YouTube et Instagram.

Le neuf partenaires du projet sCAN qui ont contribué aux exercices de monitoring sont :

- ZARA (Autriche)
- CEJI (Belgique)
- Human Rights House Zagreb (Croatie)
- Romea (Republique tchèque)
- Licra (France)
- jugendschutz.net (Allemand)
- CESIE (Italie)
- Latvian Center for Human Rights (Lettonie)
- University of Ljubljana, Faculty of Social Sciences (UL-FDV) (Slovénie)

Outre les organisations de sCAN, le secrétariat de l'INACH et les organisations partenaires du projet OpCode - ActiveWatch (Roumanie), DigiQ (Slovaquie), Estonian Human Rights Centre (Estonie), Movimiento contra la Intolerancia (Espagne) et Never Again (Pologne) - ont participé au monitoring. Pour des raisons de comparabilité, le rapport de monitoring ne comprend que les cas signalés par les partenaires du projet sCAN.

Les résultats de ce monitoring ne sont pas à considérer comme une analyse exhaustive sur l'ampleur des discours de haine sur les réseaux sociaux, mais ils représentent uniquement le contenu relevé par les organisations sur une période précise de six semaines et sur les plateformes surveillées. Certains des participants se sont concentrés sur certaines formes de discours de haine en ligne en particulier, ce qui peut avoir un impact sur les cas signalés au cours du monitoring. Ce facteur sera donc

abordé plus en détail ci-dessous. De plus, l'exercice de monitoring était centré sur la réaction des sociétés informatiques plutôt que sur le contenu spécifique des discours de haines identifiés.

Méthodologie

La méthodologie utilisée pour les exercices de monitoring respecte le processus établi par la Commission européenne au cours des précédentes périodes de monitoring. Les organisations qui ont participé ont d'abord recueilli des exemples de discours de haine sur les réseaux sociaux qui ont adhéré au code de conduite de l'UE sur la lutte contre les discours haineux illégaux en ligne. Le caractère illégal du contenu a été évalué sur la base des lois nationales qui transposent le Décision-cadre 2008/913/JAI sur la lutte contre certaines formes et manifestations de racisme et de xénophobie au moyen du droit pénal

Afin de tester la réaction des sociétés informatiques aux signalements réalisés en tant qu'utilisateurs lambda, le contenu était d'abord signalé à travers les dispositifs de signalements publics des sociétés respectives. À la suite de ces signalements, les organisations partenaires ont recensé si les sociétés informatiques avaient réagi ou pas aux signalements, en retirant ou en limitant l'accès au contenu (géo-blocage, fonctionnalités limitées, etc.) dans un délai mutuellement convenu (24h, 48h, 1 semaine). Les partenaires ont également précisé s'ils avaient reçu ou non une réponse de la part des sociétés informatiques après leur signalement, et le cas échéant dans quel délai. Fournir une réponse aux utilisateurs qui effectuent des signalements est crucial car cela permet de les impliquer et de les inciter à signaler un contenu illégal.

Certaines organisations partenaires ont réalisé une étape supplémentaire en reportant le contenu qui n'avait pas été retiré dans la semaine qui suivait le premier signalement, mais cette fois via des canaux de signalements disponibles uniquement pour les organisations partenaires, reconnues par les sociétés comme des « trusted flaggers ». Après le deuxième signalement, les organisations partenaires suivaient à nouveau le processus de monitoring et recensaient la réaction et le feedback reçu par les sociétés informatiques.

Les organisations membres de sCAN ont convenu de faire une distinction entre le contenu retiré de la plateforme et le contenu dont l'accès a uniquement été limité. Les restrictions d'accès ont, pour la quasi-totalité (99%), pris la forme du géo-blocage, rendant ainsi le contenu indisponible pour les utilisateurs qui se connectent dans le pays de signalement. Parmi les autres formes de restrictions, on retrouve la limitation de certaines fonctionnalités (comme les commentaires) sur le contenu ou le fait de le marquer comme contenu sensible. Les partenaires de sCAN considèrent que les restrictions sur le contenu ne sont efficaces qu'en partie, puisque le contenu reste en ligne et que les méthodes pour contourner les restrictions sont largement connues par les utilisateurs.

Lors du premier monitoring, les données ont été récoltées grâce à un modèle en ligne conçu et géré par la Commission européenne. Les cas ont également été enregistrés dans des fichiers Excel pour une analyse interne par les partenaires de sCAN. Lors du deuxième monitoring, les partenaires ont convenu d'utiliser un modèle Excel standardisé basé sur les suggestions des partenaires de sCAN et préparé par le secrétariat de l'INACH.

Chiffres clés

Les résultats de ce monitoring ne sont pas à considérer comme une analyse exhaustive sur l'ampleur des discours de haine sur les réseaux sociaux. Ils sont uniquement représentatifs du contenu relevé par les organisations sur une période précise de six semaines et sur les plateformes surveillées. Certains des participants se sont concentrés sur certaines formes de discours de haine en ligne en particulier, ce qui peut avoir un impact sur les cas signalés au cours du monitoring. Ce facteur sera donc abordé plus en détail ci-dessous. De plus, l'exercice de monitoring était centré sur la réaction des sociétés informatiques plutôt que sur le contenu spécifique des discours de haines identifiés.

Troisième Monitoring: 4 novembre - 13 décembre 2019

Le troisième monitoring de sCAN organisé par la Commission européenne s'est déroulé du 4 novembre au 13 décembre 2019. Pendant les six semaines de cet exercice, les partenaires du projet ont signalé 635 cas de discours haineux illégaux aux sociétés informatiques Facebook, Instagram, Twitter, YouTube, Dailymotion et Jeuxvideo. Facebook a reçu le plus grand nombre de rapports (280 cas), suivi de Twitter avec 198 cas. YouTube a reçu 102 signalements de discours haineux illégaux et Instagram en a reçu 37 de la part des partenaires de sCAN.

84 cas ont été portés à l'attention de la hiérarchie par les canaux réservés aux *trusted flaggers* des sociétés informatiques, après avoir été signalés par les utilisateurs généraux dans un délai d'une semaine après le signalement initial. Twitter a reçu 59 rapports de signalement, Facebook et Instagram ont reçu chacun 10 rapports de signalement et YouTube a reçu 5 rapports par le biais de canaux de signalement. Aucun cas n'a été transmis à Dailymotion et à Jeuxvideo.

Analyse de différentes formes de haine

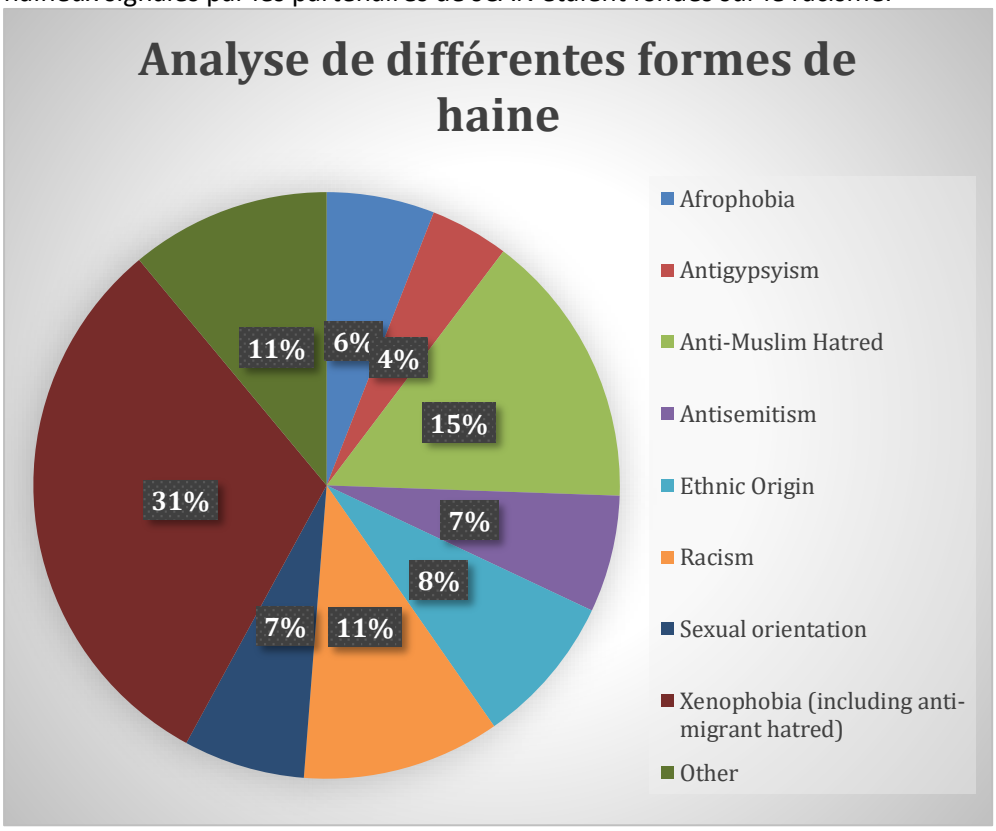
Les partenaires ont enregistré les motifs de haine qui sous-tendent le contenu des discours de haine illégaux, conformément aux catégories définies par la Commission européenne. Le type de haine le plus répandu dans l'échantillon était la xénophobie, qui comprenait celle contre les réfugiés (31% des cas). Selon l'expérience de l'INACH [1] et des projets précédemment menés par les partenaires de sCAN, la haine contre les réfugiés est devenu un phénomène répandu à partir du 2015, avec le début de la « crise des réfugiés ».[2]. Nous demandons donc instamment de séparer les chiffres relatifs à la xénophobie non liée au statut de réfugié (perçu) de la cible de la haine anti-réfugiés dans les analyses ultérieures. Le deuxième type de haine le plus répandu est l'islamophobie (15 %). Elle est souvent liée à la haine anti-réfugiés, car les auteurs de contenus hianeux ont tendance à considérer que tous les réfugiés sont musulmans et que tous les musulmans sont des réfugiés.[3] 11 % des cas de discours

[1] INACH (2016). *"Kick them back into the sea" – Online hate speech against refugees*. Disponible sir https://www.inach.net/kick-them-back-into-the-sea/ (consulté le 26.03.2020).

[2] Le terme est chargé de sens normatif, car il suggère que les réfugiés eux-mêmes sont problématiques ou que l'accueil des réfugiés en soi est essentiel. Cependant, selon nous, la « crise des réfugiés » est liée à des débats publics très conflictuels, à le « sensationnalisme » croissante des migrations et à l'atmosphère de haine envers les réfugiés.

[3] Pour en savoir plus sur les discours de haine basés sur une intersection entre la religion et l'origine ethnique, consulté l'analyse : sCAN project (2020). *Intersectional Hate Speech Online*. Available at http://scan-project.eu/wp-content/uploads/sCAN_intersectional_hate_final.pdf (consulté le 26.03.2020).

haineux signalés par les partenaires de sCAN étaient fondés sur le racisme.



Graphique 1: Analyse de différentes formes de haine; exercice de monitoring de sCAN monitoring 4 novembre– 13 décembre 2019

Taux de suppression

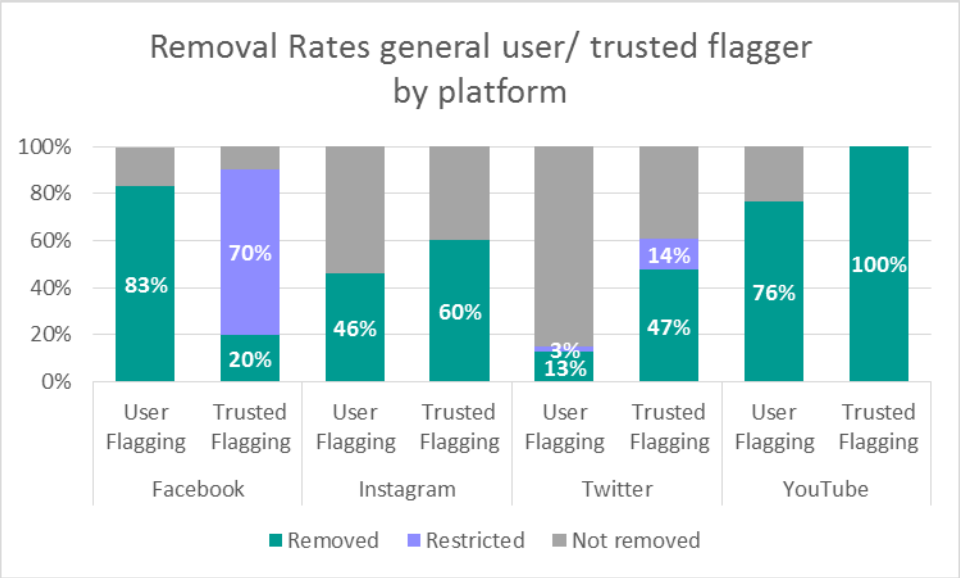
Dans l'ensemble, 67,56 % du contenu n'était plus disponible à la fin de la surveillance dans le pays d'où il provenait (64,25 % enlevé, 3,31 % restreint). Ce chiffre est conforme aux résultats des précédents exercices de monitoring menés par les partenaires de sCAN. Les sociétés informatiques ont pris des mesures dans 58,74 % des cas directement après la première notification par les canaux normaux des utilisateurs (57,80 % ont été supprimés, 0,94 % ont fait l'objet de restrictions). Certains partenaires ont fait remonter le contenu qui n'avait pas été supprimé dans la semaine suivante le premier signalement en le signalant à nouveau par les canaux disponibles pour les trusted flaggers. Les entreprises ont donné suite à 66,67 % des signalements de trusted flagger (48,81 % ont été retirés, 17,86 % ont été restreints).

Dailymotion et Jeuxvideo avaient adhéré au code de conduite peu avant l'exercice de surveillance. Comme très peu de cas leur ont été signalés (6 et 12 respectivement) et qu'ils n'ont pas été inclus dans le quatrième exercice de monitoring, pour des raisons de comparabilité, leurs chiffres sont présentés séparément. **Jeuxvideo a supprimé 100 % des cas qui lui avaient été signalés par les canaux de signalement des utilisateurs généraux dans les 24 heures.** Dailymotion a supprimé 33% des cas qui lui avaient été signalés. Les contenus supprimés, l'ont été dans les 24 heures.

Parmi les autres plateformes surveillées, Facebook a atteint le taux de retrait le plus élevé (83,21 %) pour les cas signalés par les canaux de signalement des utilisateurs généraux. YouTube a supprimé 76 % de ces cas, Instagram 46 % et Twitter n'a pris des mesures que dans 16 % des cas en supprimant 13 % et en limitant (géo-blocage) 3 % supplémentaires.

Toutes les plateformes ont obtenu de bien meilleurs résultats pour les rapports soumis par des canaux de signalement fiables. YouTube a supprimé 100 % des rapports soumis par des *trusted flaggers*. Facebook a pris des mesures dans 90 % des cas, en limitant 70 % et en supprimant 20 % des

signalements. Les partenaires du projet ne comprennent pas bien pourquoi la plateforme a choisi de restreindre un pourcentage aussi élevé de cas plutôt que de les supprimer. Instagram a supprimé 60 % des cas signalés par les trusted flaggers. L'augmentation la plus significative du taux d'action a été enregistrée pour Twitter. L'entreprise a pris des mesures dans 61 % des cas (47 % ont été supprimés, 14 % ont été restreints), soit presque **quatre fois plus** que les mesures prises dans les cas signalés par les canaux disponibles pour l'ensemble des utilisateurs.



Graphique 2: Taux de suppression par plateforme; exercice de monitoring de sCAN – 4 novembre – 13 décembre 2019

Délai de suppression

En ce qui concerne les délais d'éloignement, Facebook a une fois de plus affiché les meilleures performances. Facebook est la seule plateforme qui a pris des mesures sur la majorité des contenus signalés par les canaux d'utilisateurs généraux dans les 24 heures (76,43 %). YouTube a retiré 47,06 % et Instagram 40,54 % dans les 24 heures. Twitter n'a pris des mesures que sur 8,59 % des contenus signalés par les canaux accessibles au public dans ce délai.

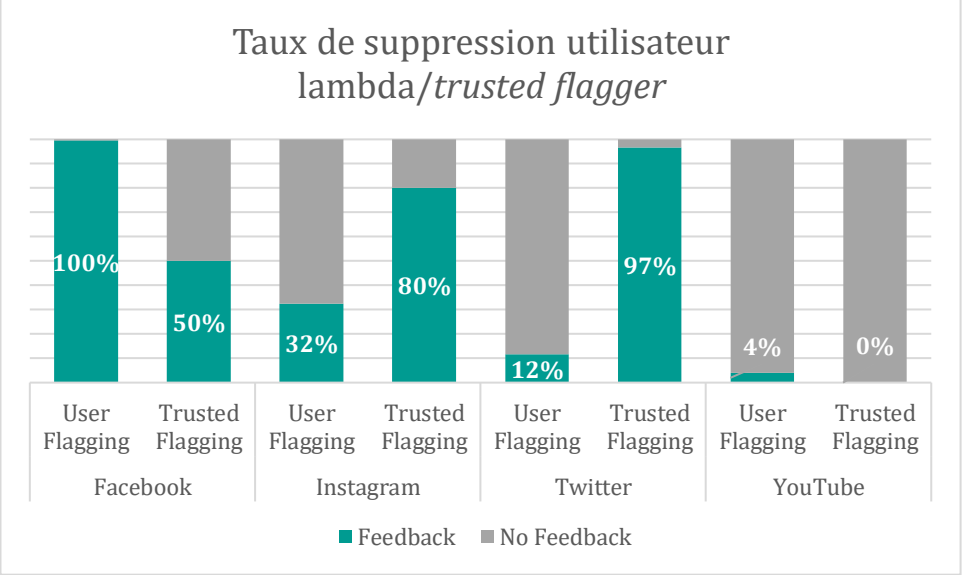
La plupart des entreprises ont réagi plus rapidement aux contenus signalés par des canaux de signalement fiables. Twitter a notamment amélioré ses performances en supprimant 47,46 % des contenus signalés par des *trusted flaggers* dans les 24 heures. Facebook a pris des mesures pour 70 % de ces contenus dans les 24 heures. Instagram a retiré 60 % et YouTube 40 % des contenus signalés par les trusted flaggers dans ce délai.

Feedback

Il est crucial pour les utilisateurs et les personnes qui dénoncent les discours haineux sur les réseaux sociaux de recevoir un retour d'information afin d'établir une relation de confiance et de leur permettre de mieux comprendre comment les plateformes modèrent le contenu. Le code de conduite signé en 2016 exige également des entreprises qu'elles fournissent un retour d'information en temps utile. Il existe une énorme différence entre les entreprises impliquées dans le code de conduite en ce qui concerne le retour d'information. L'un des principaux objectifs de l'organisation continue d'exercices de monitoring et le but du présent rapport est de fournir des informations publiques sur le retour d'information concernant les discours de haine signalés sur les réseaux sociaux.

Dans l'ensemble, les sociétés informatiques ont fourni un retour d'information à 36,9 % des rapports par les canaux disponibles pour les utilisateurs généraux et à 56,7 % des rapports par les canaux de confiance. Conformément aux exercices de monitoring précédents, le taux de retour d'information

des *trusted flaggers* est plus élevé que celui des utilisateurs normaux - surtout en moins de 24 heures (près de 20 points de différence). Il est important de noter que la situation entre les sociétés informatiques est très disparate en ce qui concerne le taux de retour d'information.



Graphique 1: Taux de suppression par plateforme; exercice de monitoring de sCAN – 4 novembre – 13 décembre 2019

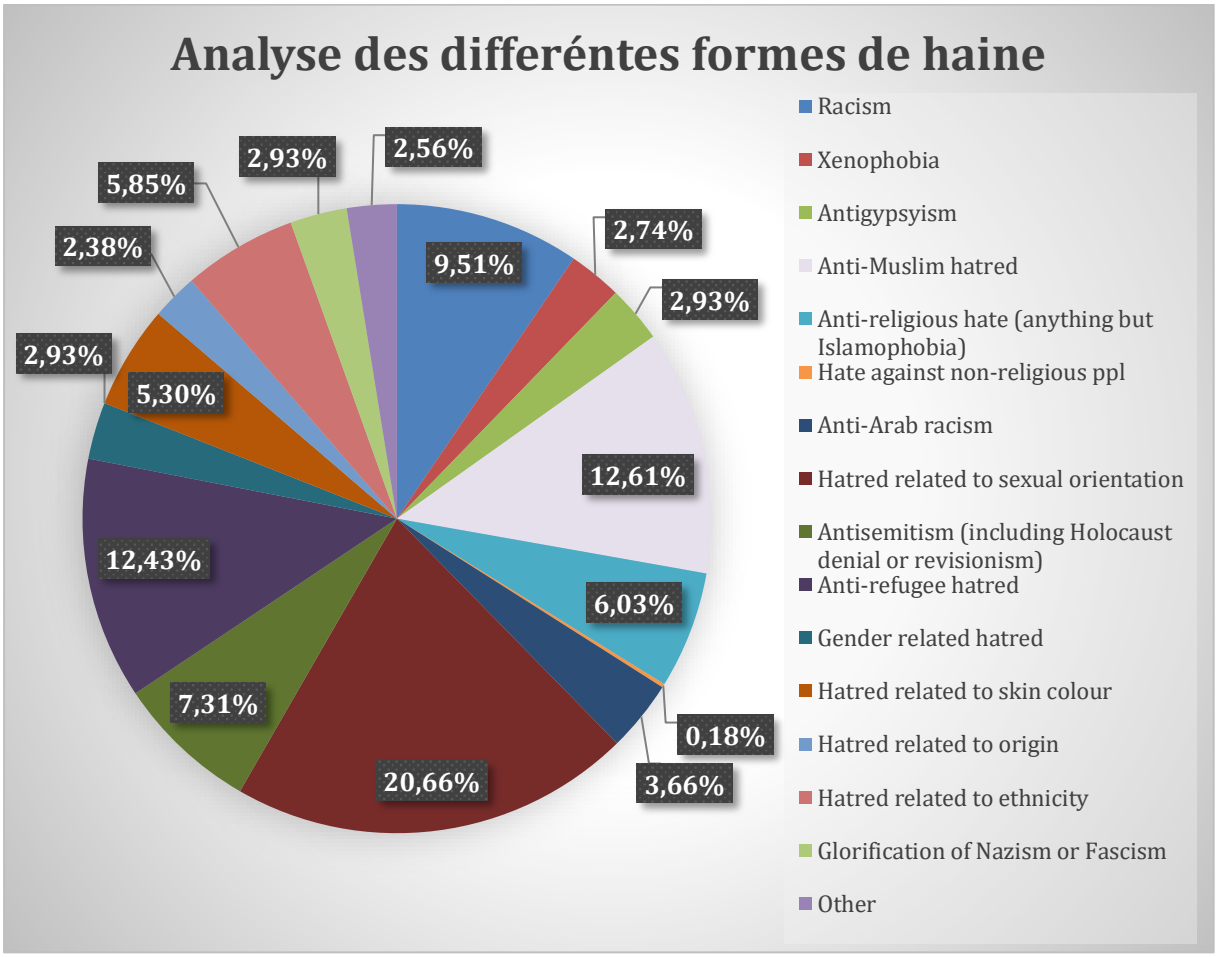
Facebook peut être considéré comme un cas isolé : les utilisateurs normaux ont presque toujours reçu un retour d'information après avoir signalé leur problème. Il s'agit là encore de la seule société informatique à fournir systématiquement un retour d'information à tous ses utilisateurs. Néanmoins, ***les trusted flaggers*** n'ont reçu un retour d'information spécifique que dans 50 % des cas.
Sur Instagram et Twitter, la priorité a été donnée au canal de signalement des trusted flaggers en ce qui concerne leur taux de retour (80 % et 96,6 %). Il semble que leurs politiques tendent à accorder plus d'attention aux *trusted flaggers*. D'autre part, YouTube a rarement, voire jamais, fourni un retour d'information sur les notifications normales et les signalements de confiance : aucun retour d'information n'a été reçu pour tous les cas signalés par les partenaires ***de sCAN*** via les canaux de signalement des signalements de confiance.
Lorsque les sociétés informatiques ont donné un retour d'information, elles l'ont presque toutes fait dans les 24 heures. Dans l'ensemble, un retour d'information a été fourni dans 91,2 % des cas signalés par les canaux des utilisateurs généraux en moins de 24 heures. Néanmoins, cette durée pourrait être prolongée pour ***les trusted flaggers*** : seulement 75 % des informations ont été fournies en moins de 24 heures.

Quatrième monitoring: 20 janvier – 28 février 2020

Le quatrième monitoring de sCAN a eu lieu entre le 20 janvier 2020 et le 28 février 2020. Il s'agissait d'une surveillance silencieuse en coopération avec le secrétariat de l'INACH et le projet OpCode. Les partenaires ont signalé 484 cas de discours haineux illégaux en ligne aux sociétés informatiques Facebook (242 cas), Twitter (127), YouTube (66) et Instagram (49). Afin de tester la réaction des sociétés informatiques aux notifications de leur base d'utilisateurs générale, les notifications ont d'abord été envoyées anonymement par des canaux accessibles au public. Dans un deuxième temps, 94 cas qui n'avaient pas été retirés après la notification en tant qu'utilisateurs généraux ont été signalés à nouveau par des canaux de notification disponibles uniquement pour les trusted flaggers.

Analyse de différentes formes de haine

Afin de fournir une image aussi complète et approfondie de la haine en ligne au sein de l'Union européenne, le projet sCAN et INACH ont classé les cas de discours haineux en ligne au cours de cet exercice de surveillance en quinze catégories différentes. Une catégorie « autre » a également été ajoutée afin de pouvoir enregistrer des cas qui, autrement, passeraient entre les mailles du filet. Par exemple, le partenaire français Licra a signalé des cas de discours de haine anti-asiatique liés à l'épidémie de Covid-19, qui ont déjà fait l'objet d'une cehrtaine attention en février 2020.



Graphique 4: Analyse de différentes formes de haine; exercice de monitoring de sCAN monitoring 20 janvier – 13 février 2020

La haine liée à l'orientation sexuelle était le type de haine le plus répandu dans l'échantillon de cas collectés au cours de cet exercice de monitoring. Cela peut être lié à des causes et à des événements locaux spécifiques qui ont influencé les débats publics dans certains pays suivis pendant la période de monitoring. La plupart des cas homophobes ont été enregistrés en Croatie et en Lettonie. La Human Rights House Zagreb (HRHZ) a recueilli au total 49 cas, dont 48 étaient homophobes et le Latvian Centre for Human Rights (LCHR) a recueilli près de 50 cas homophobes, tout en recueillant près du double du nombre de cas que toute autre organisation participant à l'exercice de monitoring. Le HRHZ et le LCHR ont tous les deux signalé que, pendant leur période de collecte de données, de nombreux reportages ont placé les questions LGBTQ+ au centre du débat public. En Croatie, un débat public est en cours sur la question de savoir si les couples homosexuels seraient autorisés à être parents d'accueil. En Lettonie, de nombreux articles ont été publiés sur un homme politique letton qui a épousé son partenaire de même sexe en Allemagne et sur un joueur de basket letton qui a eu un enfant avec son partenaire de même sexe. Ces événements ont été à l'origine de discours de haine envers la communauté LGBTQ+ dans ces pays.

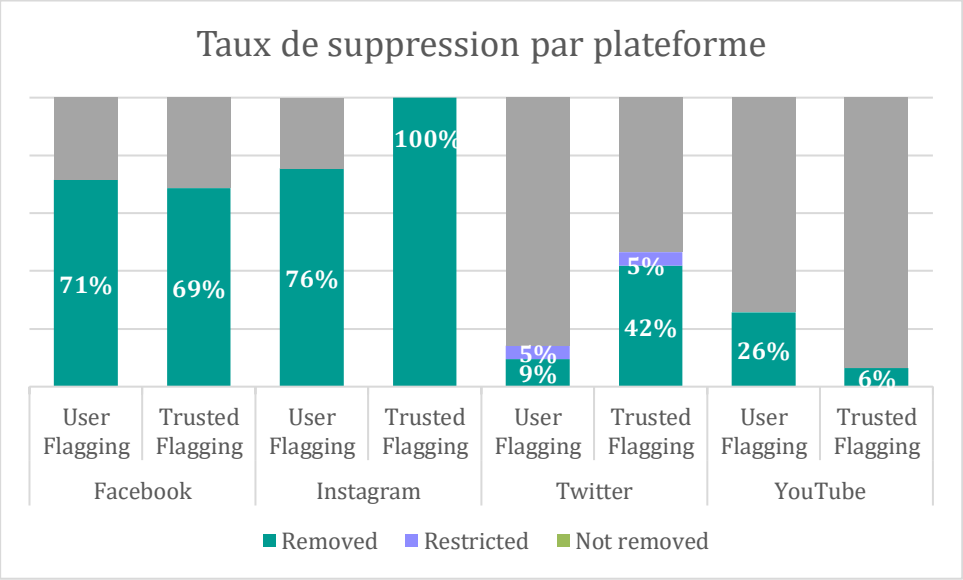
Hormis cette particularité, les types de haine représentés dans cet échantillon sont conformes aux conclusions des précédents exercices de monitoring. L'islamophobie (12,61 %) et la haine anti-réfugiés (12,43 %) arrivent en tête de l'échelle, tête à tête. Un résultat peu surprenant, puisque les deux sont intimement liés. Ils sont suivis par le racisme (9,51 %) et l'antisémitisme (7,31 %). Les discours de haine antisémites, y compris la négation de l'Holocauste, étaient particulièrement répandus à l'occasion de la Journée de commémoration de l'Holocauste le 27 janvier.

Taux de suppression

Dans l'ensemble, seuls 58 % des cas signalés n'étaient plus disponibles à la fin du monitoring. Il s'agit d'une baisse importante par rapport au troisième exercice de monitoring effectué seulement un mois plus tôt. Cela souligne l'importance d'un traitement cohérent des cas par les plateformes, indépendamment des exercices de monitoring officiels organisés par la Commission européenne.

51 % des cas ont déjà été retirés après les notifications initiales en tant qu'utilisateurs généraux (signalement normal de l'utilisateur). Instagram a atteint le taux de suppression le plus élevé avec 75,51 % des cas supprimés après la notification par les canaux des utilisateurs généraux. Facebook a supprimé 71,49 % des cas après la notification initiale. YouTube et Twitter ont obtenu des résultats nettement moins bons. YouTube a supprimé 25,76 % des cas après notification par l'utilisateur, tandis que Twitter n'en a supprimé que 9,45 % et a restreint 4,72 % de ces cas.

94 cas ont été transmis par des canaux de signalement fiables après ne pas avoir été supprimés par les entreprises lorsqu'ils ont été signalés par les canaux de notification générale des utilisateurs. Parmi ces cas, 39 % ont été supprimés par les sociétés informatiques. Instagram a supprimé tous les cas qui lui avaient été signalés une deuxième fois par les canaux de signalement de confiance. Facebook a supprimé 68,75 % des cas signalés par les trusted flaggers. Twitter a supprimé un pourcentage beaucoup plus élevé de cas lorsqu'ils étaient signalés par des canaux de signalement de confiance (41,86 %) et en a restreint encore 4,65 %, tandis que YouTube a supprimé moins de cas (6,45 %) que lorsqu'ils étaient signalés par des utilisateurs généraux.



Graphique 2: Taux de suppression par plateforme; exercice de monitoring de sCAN –20 janvier – 28 février 2020

Évaluation et délai de suppression

Dans le code de conduite, les plateformes s'engagent à évaluer et à supprimer la majorité des cas illégaux qui leur sont signalés en moins de 24 heures. Les partenaires de sCAN ont compté le retrait des cas et/ou ont fourni un retour d'information comme évaluation. Dans le cadre de cet exercice de

monitoring, les plateformes ont atteint cet objectif dans 44,21 % des cas signalés en tant qu'utilisateurs normaux. 4,13 % des cas ont été évalués après 48 heures et 4,96 % ont été évalués en une semaine. Dans 46,69 % des cas, il n'y avait aucune indication d'une évaluation une semaine après le rapport initial.

Seules deux plateformes ont retiré la majorité des contenus signalés par les canaux de signalement généraux dans les 24 heures. Instagram a retiré 71,43 % des contenus signalés dans les 24 heures, sa société mère Facebook a retiré 59 % des contenus signalés dans ce délai. Twitter (2,36 %) et YouTube (1,52 %) n'ont pratiquement retiré aucun contenu signalé en tant qu'utilisateur général dans les 24 heures. Lorsque les partenaires ont signalé du contenu par l'intermédiaire de leurs canaux de signalement de confiance, davantage de contenu a été retiré dans les 24 heures. La plus grande différence de temps de retrait entre le signalement par les utilisateurs normaux et le signalement par les trusted flaggers a été signalée pour Twitter, qui a retiré 37,21 % des contenus signalés par les trusted flaggers dans les 24 heures. Instagram (75 %) et Facebook (62,5 %) ont encore amélioré leur performance lorsque le contenu était signalé par des *trusted flaggers*. YouTube a retiré 6,45 % des contenus signalés par des *trusted flaggers* dans les 24 heures.

Près de trois quarts des cas évalués par les entreprises dans la semaine suivant la notification ont été retirés par les sociétés informatiques, qui ont également fourni un retour d'information au partenaire à l'origine du signalement. 10 % des cas évalués ont été supprimés, mais l'organisation qui a fait la déclaration n'a pas reçu de retour d'information de la part de l'entreprise. Dans 17 % des cas évalués, l'entreprise a fourni un retour d'information informant l'organisme déclarant que le contenu était jugé non conforme aux normes communautaires et n'a donc pas été supprimé.

Dans les cas où il n'y avait pas d'indication d'évaluation après une semaine, les partenaires ont vérifié à la fin du monitoring si les cas étaient toujours en ligne. La grande majorité (86,36 %) de ces cas étaient toujours en ligne après la fin du monitoring et les partenaires n'ont reçu aucun retour sur leur notification.

Dans 4,09 % des cas, les partenaires ont reçu un retour d'information plus d'une semaine après leur notification pour les informer que le contenu avait été déplacé. Dans 1,36 % des cas, le contenu n'a pas été retiré, mais les partenaires ont reçu un retour d'information les informant de cette décision plus d'une semaine après leur rapport aux sociétés informatiques. 8,18 % des cas ont été supprimés à un moment donné entre une semaine après la notification et la fin du contrôle, mais les sociétés informatiques n'ont pas informé les organismes déclarants de la suppression. Il est donc impossible de dire si les cas ont été supprimés à la suite du monitoring ou pour d'autres raisons.

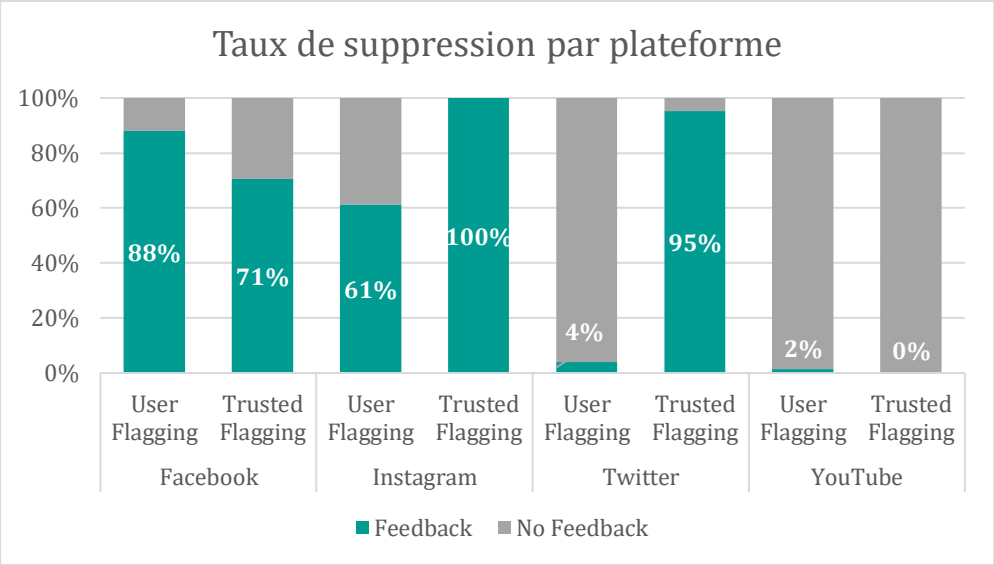
Dans les cas où l'information avait été remontée ils ont été pour la plupart évalués dans les 24 heures suivant le rapport (52 %). 2 % ont été évalués dans les 48 heures et 5 % dans la semaine. Il n'y a eu aucune indication d'évaluation dans 41 % des cas. Les sociétés informatiques n'ont pas supprimé ces cas et n'ont pas fourni de retour d'information aux organisations qui ont fait rapport, même si ces organisations sont enregistrées comme des trusted flaggers.

Les plateformes ont eu des résultats différents lorsqu'elles ont évalué les cas qui leur ont été signalés par des trusted flaggers. Twitter (83,72 %), Instagram (75 %) et Facebook (62,5 %) ont évalué la majorité de ces cas en moins de 24 heures. Cependant, rien n'indique qu'il y ait eu une évaluation (soit un retour d'information, soit un retrait) des cas signalés à YouTube en tant que trusted flaggers. Il est essentiel de recevoir un retour d'information sur les discours de haine signalés (même s'ils ne sont pas supprimés par la plateforme) pour que la coopération entre les plateformes de réseaux sociaux et leurs trusted flaggers soit fructueuse. Il serait donc très apprécié de recevoir davantage de communications de YouTube sur les cas signalés afin d'améliorer la coopération

Feedback

Les sociétés informatiques ont fourni un retour d'information pour 51,45 % des rapports par les canaux disponibles aux utilisateurs normaux (42,56 % en moins de 24 heures) et pour 60 % des rapports par les canaux de confiance (52,63 % en moins de 24 heures). Selon l'analyse des exercices de monitoring précédents, le taux de retour d'information des trusted flaggers est plus élevé que celui des utilisateurs normaux - surtout en moins de 24 heures : près de 10 points de différence pour un retour d'information en moins de 24 heures. Toutefois, par rapport au dernier exercice de monitoring, l'écart a été réduit de moitié.

Facebook a fourni moins de retours d'information aux utilisateurs par rapport au dernier exercice de monitoring - environ 88 %. Mais, pour cet exercice, le taux de retour des trusted flaggers a augmenté à environ 70,6 %.
Les taux de retour d'information d'Instagram se sont considérablement améliorés : ils sont presque doublé pour les utilisateurs (de 32 % pour le 3e exercice de monitoring à 61 % pour ce 4e exercice de monitoring) et le signaleur de confiance a reçu un retour d'information dans 100 % des cas.
Les taux de retour de Twitter et YouTube sont restés faibles. Le taux de retour général des utilisateurs de Twitter s'est détérioré, passant à 11,6 % pour ce 4e exercice de monitoring. Le taux de retour de la plateforme a été de 96,6 %.



Graphique 3: Taux de suppression par plateforme; exercice de monitoring de sCAN –20 janvier – 28 février 2020

Pour YouTube, la situation est fondamentalement la même pour les rapports effectués par les canaux normaux d'utilisateurs et par les canaux de signalement de confiance et en cohérence avec le 3ème exercice de surveillance : YouTube n'a envoyé aucun retour d'information sur les cas signalés par les partenaires de sCAN par l'intermédiaire de canaux de signalement fiables.

Dans ce quatrième exercice de monitoring, Facebook et Instagram ont mis un peu plus de temps à fournir un retour d'information. Facebook a fourni un retour d'information en moins de 24 heures pour 71,9 % des cas signalés par les comptes d'utilisateurs généraux. Sur Instagram, le taux est de 61,2 %.

Il est également important de souligner que le temps de retour d'information sur Twitter a considérablement augmenté : lors de notre 3ème exercice de monitoring, dans 82 % des cas, le retour

d'information a été fourni aux utilisateurs normaux dans les 24 heures. Cependant, cela n'a été que dans 1,57 % des cas lors de notre quatrième exercice de monitoring.

Près de la moitié des plateformes ont envoyé plus de commentaires aux trusted flaggers qu'aux rapports généraux des utilisateurs. Pour Facebook, comme dans les exercices précédents, le taux de retour est plus important pour les rapports des utilisateurs généraux que pour les rapports des trusted flaggers. En ce qui concerne YouTube, comme déjà mentionné, les deux taux sont très faibles.

Expériences et observations

Afin de bénéficier des expériences et des observations des partenaires pour les futurs exercices de monitoring, les partenaires de sCAN ont rempli un questionnaire d'évaluation à la fin du monitoring. Certaines observations importantes seront examinées dans cette section.

Les partenaires ont indiqué que le dispositif utilisé pour l'établissement des rapports semblait avoir un impact sur le fait qu'ils reçoivent ou pas un retour d'information de la part d'Instagram. Alors que les partenaires qui ont fait leurs rapports par l'intermédiaire de l'application mobile ont déclaré avoir reçu un retour d'information de la plateforme, les partenaires qui ont fait leurs rapports à Instagram en utilisant un ordinateur de bureau n'ont pratiquement pas reçu de retour d'information.

Au cours de la période de monitoring, les partenaires ont remarqué que plusieurs comptes affichaient un grand nombre de commentaires et de messages haineux illégaux. Certains de ces pages ou comptes ont publié quotidiennement un nombre important de commentaires racistes, misogynes et extrêmement violents. Par conséquent, nous recommandons aux sociétés informatiques de surveiller ces comptes de plus près et de prendre des mesures décisives contre chaque cas de discours haineux illégal publié sur ces comptes.

En répondant à la question de savoir comment l'exercice de surveillance a eu un impact sur les personnes qui rapportent les contenus, les partenaires ont souligné l'impact de l'exercice de surveillance sur leur travail, qui prend beaucoup de temps. Certains ont résolu le problème en répartissant la charge de travail entre différents chercheurs. Il y avait aussi le problème de devoir traiter un nombre croissant de cas signalés dans le cadre du travail régulier de l'organisation et donc une capacité réduite à faire remonter les cas liés à l'exercice de monitoring.

Pour certaines organisations ayant une grande expertise dans la conduite d'exercices de monitoring, il s'agissait simplement d'un monitoring de plus et il n'y avait pas d'impact supplémentaire sur les experts, les assistants ou les chercheurs.

En outre, certains partenaires ont souligné une série de facteurs psychologiques liés au travail de monitoring. Le chercheur de ROMEA, par exemple, a indiqué que l'exercice de monitoring avait suscité des sentiments de « dégoût et d'impuissance », mais aussi une détermination à poursuivre.

Il est particulièrement intéressant de noter que certaines organisations offrent une forme de conseil en cas de traumatisme ou de détresse liée au monitoring. Dans le cas du CESIE, cela s'est fait en discutant au sein de l'équipe de la manière dont l'humeur des participants a été affectée par le contenu auquel ils ont été exposés. Ils ont également signalé qu'une de leurs chercheuses était devenue hyperactive en continuant à rapporter les contenus même au-delà de ses heures de travail et de sa charge de travail. Dans le cas de jugenschutz.net, les employés ont régulièrement accès à des conseils en matière de traumatisme.

En conclusion, nous pouvons dire que les exercices de monitoring ont des effets psychologiquement et émotionnellement éprouvants sur les rapporteurs, qu'ils prennent du temps, qu'ils doivent être partagés de préférence au sein de l'organisation par quelques rapporteurs et qu'il est préférable de disposer d'une forme de conseil en matière de traumatisme. La synchronisation du moment de

l'exercice de monitoring avec le rythme de l'organisation peut également jouer un rôle dans la réussite du monitoring

Conclusion

Les résultats de ces exercices de monitoring mettent en évidence la nécessité d'une performance plus cohérente des sociétés informatiques dans la suppression des discours de haine illégaux en ligne. Le taux global d'élimination de 58% lors du quatrième monitoring est inférieur de près de 10 points de pourcentage au taux global d'élimination des exercices précédents. Cela inclut le troisième exercice de surveillance de sCAN en novembre et décembre 2019, un mois seulement avant. Les entreprises doivent à tout moment s'assurer qu'elles répondent en temps utile et suppriment les discours haineux illégaux en ligne.

La plupart des entreprises fournissent davantage de commentaires aux auteurs des signalements de confiance qu'à leur base d'utilisateurs traditionnels. Cela peut être problématique, car les organisations de la société civile reconnues comme *trusted flaggers* ne peuvent pas surveiller et signaler tous les discours de haine illégaux par elles-mêmes.

Il est essentiel d'impliquer tous les utilisateurs des plateformes dans le signalement des discours de haine pour lutter efficacement contre les discours de haine illégaux en ligne. Le retour d'information est un aspect important pour maintenir l'engagement et la motivation des utilisateurs à signaler, ainsi que pour leur permettre de mieux comprendre comment les plateformes modèrent le contenu et appliquent les normes issues de l’Union européenne.

Bibliographie

European Union (2008). *COUNCIL FRAMEWORK DECISION 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law*. Disponible sur https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008F0913&from=EN (consulté le 26.03.2020).

INACH (2016). *"Kick them back into the sea" – Online hate speech against refugees*. Disponible sur https://www.inach.net/kick-them-back-into-the-sea/ (consulté le 26.03.2020).

sCAN project (2020). *Intersectional Hate Speech Online*. Disponible sur http://scan-project.eu/wp-content/uploads/sCAN_intersectional_hate_final.pdf (consulté le 26.03.2020).