Platforms, Experts, Tools: Specialised Cyber-Activists Network

Rapport sur les exercices de monitoring

2018 – 2019

Projet financé par le programme « droits, égalité et citoyenneté » (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 ».

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

Sommaire

Introduction

Au cours de la première année du projet, les organisations partenaires ont participé à deux exercices de monitoring conjoints avec la Commission européenne et l'INACH (International Network Against Cyber Hate). L'objectif des exercices de monitoring était d'évaluer dans quelle mesure Facebook, Twitter, Youtube, et Instagram observent le Code de conduite visant à combattre les discours de haine illégaux en ligne, élaboré en 2016 par la Commission européenne. De 2016 à 2018, la Commission européenne a organisé quatre périodes de monitoring visant à évaluer le Code de conduite. La plupart des partenaires sCAN avait déjà participé aux précédents exercices de monitoring organisés par la Commission et par l'INACH.

Dans le Code de conduite, les sociétés informatiques s'engagent à « examin[er] la majorité des signalements valides en moins de 24 heures »[1] et à retirer ou limiter l'accès au contenu qui enfreint leurs orientations communautaires et/ou la loi du pays. Puisqu'il est impossible pour des organisations externes d'évaluer le délai d'examen d'un signalement, les partenaires sCAN ont enregistré le délai qu'il a fallu aux sociétés pour agir ou donner un retour au sujet des signalements.

Le premier monitoring qui a eu lieu au cours du projet sCAN a été organisé par la Commission européenne et s'est déroulé du 5 novembre au 14 décembre 2018. Au cours de cette période, les partenaires sCAN ont signalé 762 cas de discours de haine illégaux en ligne aux sociétés informatiques suivantes : Facebook, Twitter, YouTube, Instagram, Google+, Dailymotion et Jeuxvidéo.

Le deuxième monitoring a été conjointement organisé par les partenaires sCAN et par l'INACH et s'est déroulé du 6 mai au 21 juin 2019. Les partenaires ont alors signalé 432 cas aux sociétés informatiques suivantes : Facebook, Twitter, YouTube, et Instagram.

Les neuf partenaires sCAN suivants ont participé aux exercices de monitoring :

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

En plus de ces derniers, le secrétariat INACH et des organisations partenaires du réseau, à savoir Greek Helsinki Monitor (Grèce) et Never Again Association (Pologne), ont participé au deuxième monitoring.

Les résultats de ce monitoring ne sont pas à prendre 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 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.

[1] Commission européenne (2016). Code de conduite visant à combattre les discours de haine illégaux en ligne. Disponible sur https://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=54300 (consulté le 28.08.2019).

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 participés ont d'abord recueilli des exemples de discours de haine sur les réseaux sociaux étudiés. 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[2].

Afin de tester la réaction des sociétés informatiques aux signalements réalisés en tant qu'utilisateurs lambdas, le contenu était d'abord signalé à travers les dispositifs de signalement publics des sociétés respectives. À la suite de ces signalements, les organisations partenaires ont recensé si oui ou non les sociétés informatiques avaient réagi 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 recensé s'ils avaient reçu ou non une réponse de la part des sociétés informatiques après leur signalement, et dans quels délais. Fournir une réponse aux utilisateurs qui effectuent des signalements est crucial car cela permet de les impliquer et de les motiver 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 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 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 des utilisateurs.

Afin de permettre une analyse conjointe et une comparaison des résultats, les organisations qui ont participé ont recensé les cas relevés dans des bases de données transnationales. Pour le premier monitoring, les données ont été récoltées grâce à un modèle en ligne conçu et géré par la Commission européenne. Pour le deuxième monitoring, les partenaires se sont mis d'accord pour utiliser la base de données de l'INACH sur les discours de haine. La base de données de l'INACH a été réalisée dans l'optique de créer un instrument international permettant de documenter et d'analyser les cas de haine en ligne, et pour servir de point de contact central pour signaler les cas de haine en ligne.

[2] Union européenne (2008). DÉCISION-CADRE 2008/913/JHA sur la lutte contre certaines formes et manifestations de racisme et de xénophobie au moyen du droit pénal. Disponible à l'adresse : https://eur-lex.europa.eu/eli/dec_framw/2008/913/oj?locale=fr (consulté le 28.08.2019).

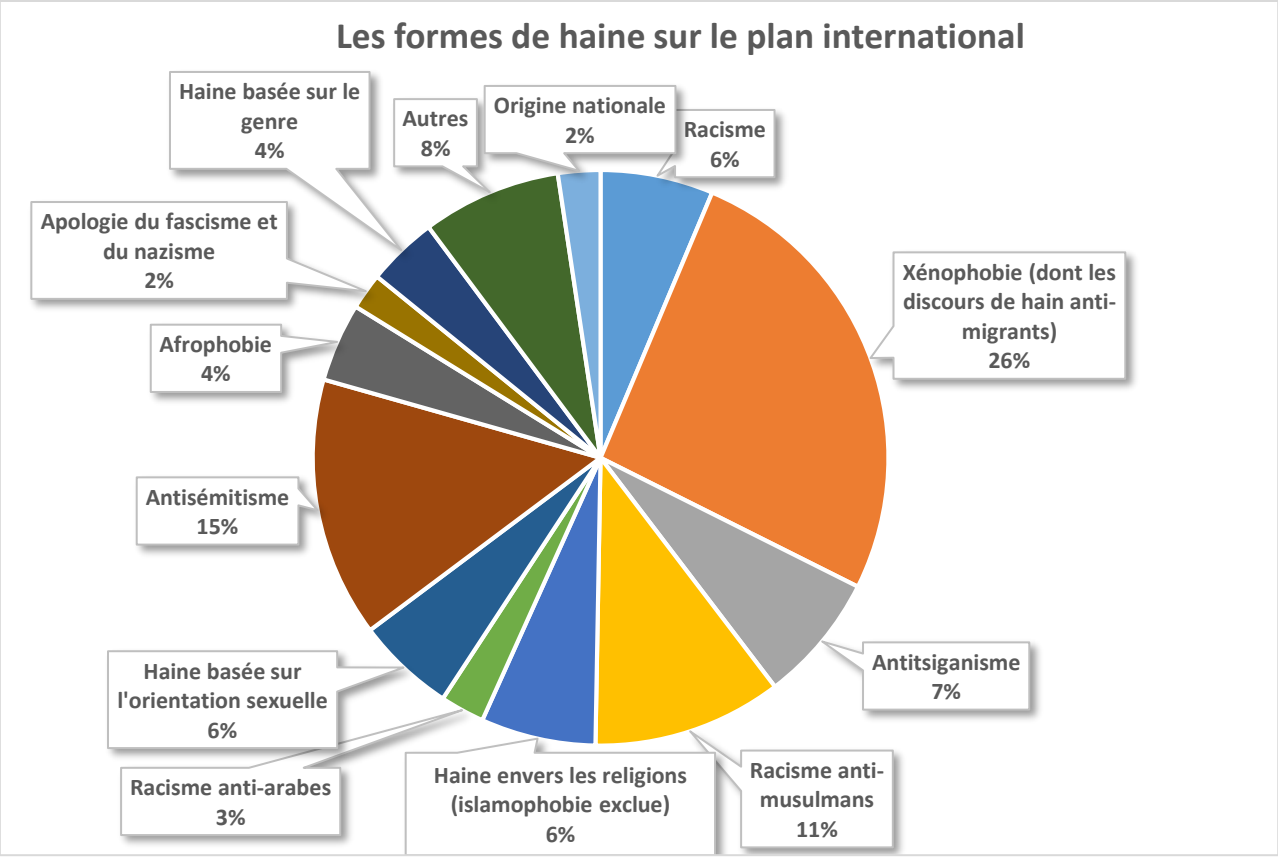
Chiffres clés

Les résultats de ce monitoring ne sont pas à prendre 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 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.

Premier monitoring : 5 novembre –14 décembre 2018

Le premier monitoring s'est déroulé du 5.11.2018 au 14.12.2018. Au cours de cette période, les partenaires sCAN ont signalé 762 cas de contenus haineux illégaux en ligne à Facebook (311 cas), Twitter (190), YouTube (142), Instagram (86), Google+ (23), Dailymotion (8) et Jeuxvidéo (2). De façon à tester les réactions des sociétés informatiques face aux signalements effectués par leurs utilisateurs lambdas, 755 signalements ont été envoyés de façon anonyme via des dispositifs disponibles au public. Dans un second temps, 165 cas qui n'avaient pas été retirés après signalement par des « utilisateurs lambdas » ont été signalé à nouveau via des dispositifs de signalements disponibles uniquement pour les trusted flaggers. Sept cas ont également été signalés directement via ces trusted flaggers, ce qui fait un total de 172 signalements.

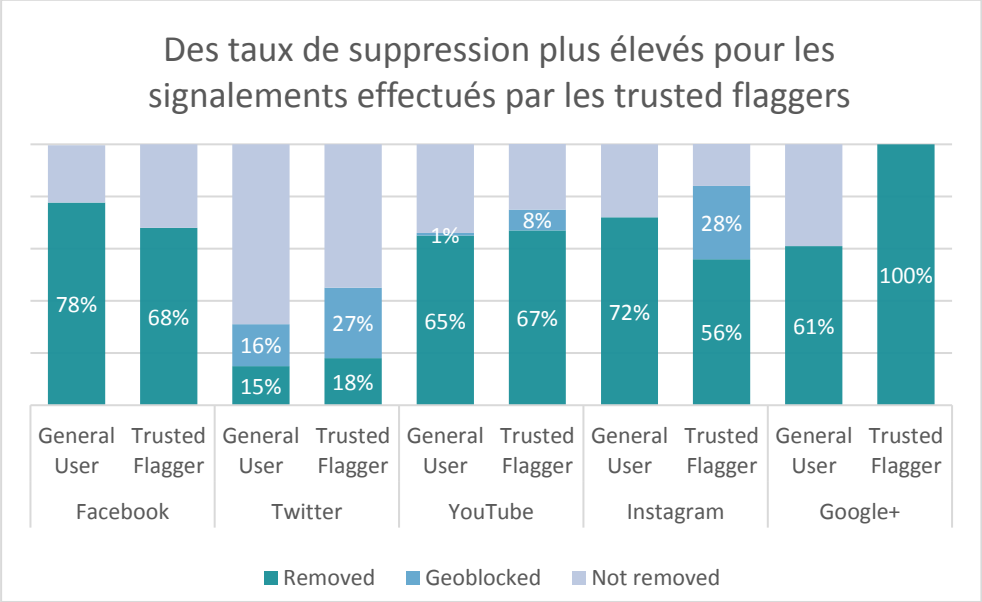
Analyse des différentes formes de haine



Graphique 1 : Les formes de haine sur le plan international ; Source : monitoring sCAN

Pour le premier monitoring, des informations ont été recensées au sujet des formes de haine correspondant aux catégories établies par la Commission européenne pour toutes les organisations participant au monitoring. Les formes de haine les plus courantes dans l'échantillon relevé par les partenaires du projet sCAN étaient la xénophobie (dont les discours de haine envers les migrants) (30%), l'antisémitisme (17%), et le racisme anti-musulmans (12%).

Taux de suppression :



Graphique 2: Taux de suppression utilisateur lambda/trusted flagger ; Source : sCAN monitoring

Au total, les sociétés surveillées ont agi dans 73% des cas, en retirant (67%) ou en géo-bloquant (6%) le contenu. Les taux de suppression différaient en fonction du moyen de signalement utilisé. Au total, les sociétés informatiques ont réagi dans 62% des cas pour le contenu signalé en tant qu'utilisateurs lambdas (58% de suppression, 4% de géo-blocage) et dans 60% des cas pour le contenu signalé en tant que trusted flaggers (42% de suppression, 18% de géo-blocage). La plupart des sociétés informatiques ont réagi plus souvent face aux signalements effectués via les dispositifs privés que face à ceux effectués via les dispositifs disponibles publiquement aux utilisateurs de leurs plateformes.

Les taux de suppression ne différaient pas seulement en fonction des modes de signalements utilisés. En effet, les partenaires sCAN ont également observé **des différences entre les pays dans la façon de répondre aux signalements des trusted flaggers**. Facebook a retiré 100% des cas signalés par les trusted flaggers en Lettonie, en Allemagne, en France, et en République Tchèque, mais seulement 50% en Autriche[3]. Twitter et Instagram n'ont géobloqué que les signalements effectués par les trusted flagger français.

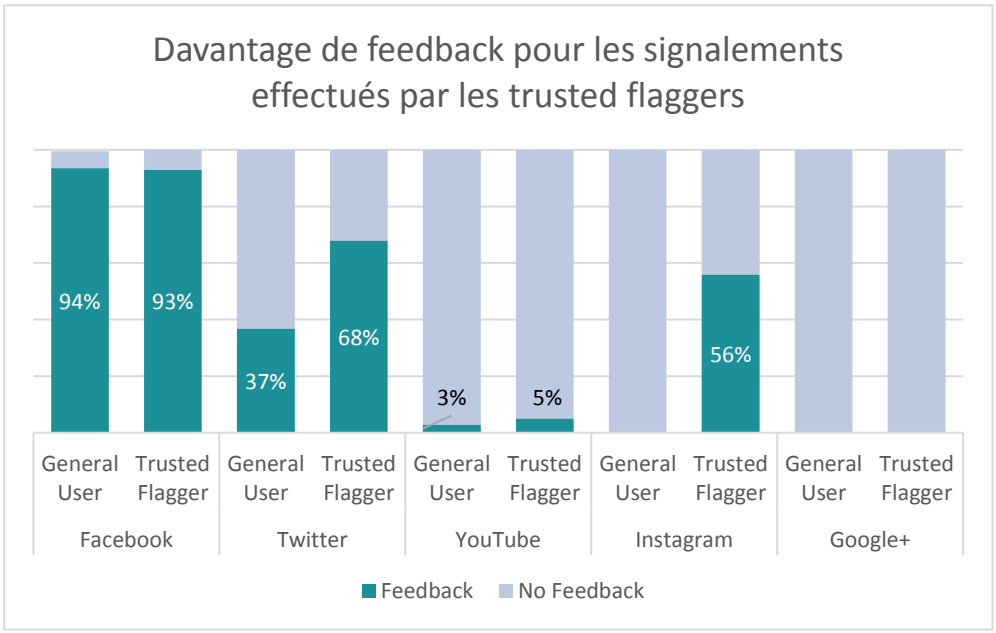
[3] Les contenus signalés par le partenaire autrichien en tant que trusted flaggers qui n'ont pas été supprimés par Facebook ont été évalués par des experts juridiques qui les ont majoritairement estimés pertinents selon le droit pénal autrichien, mais ils sont aussi extrêmement complexes. Juger ces cas illégaux et non respectueux des standards de la communauté de Facebook nécessite une connaissance approfondie du contexte. Il est alors probable que Facebook n'ait pas été capable de saisir la complexité (juridique) des cas signalés, pour cause d'un manque de connaissance du contexte national et de la situation politique et de la façon d'appréhender les différents dialectes allemands. Facebook n'a pas saisi cette opportunité pour consulter le trusted flagger.

Délai de suppression :

Puisqu'il est impossible pour des organisations externes de mesurer la durée d'examen d'un signalement, les partenaires sCAN ont enregistré le temps que les entreprises notifiées ont mis à réagir ou à fournir un feedback. Deux des sociétés informatiques suivies ont retiré la majorité du contenu moins de 24 heures après avoir reçu le signalement à travers les dispositifs de signalement publics : Facebook (76%) et YouTube (58%). Instagram a retiré 47% du contenu en moins de 24 heures et Google+ 35%. Twitter a retiré 12% du contenu en 24 heures et en a géo-bloqué 13%. Lorsque les signalements ont été effectué via les dispositifs de signalement privés, YouTube a retiré le contenu dans 67% des cas et l'a géo-bloqué dans 8% en moins de 24 heures ; Instagram en a retiré 50% et géo-bloqué 28%, Twitter en a retiré 17% et géo-bloqué 27%, et Facebook en a retiré 32%. Google+ n'a retiré aucun contenu signalé par les trusted flaggers en moins de 24 heures.

Feedback :

Au total, les sociétés informatiques ont fourni une réponse pour 48% des signalements effectués via les dispositifs publics (46% en moins de 24 heures) et pour 55% de ceux effectués via les dispositifs privés des partenaires (45% en moins de 24 heures). Facebook est la seule société informatique qui a fourni une réponse à chaque fois, à tous les utilisateurs, alors que Twitter et YouTube ont plus souvent fourni un feedback pour les signalements effectués par des trusted flaggers que pour ceux effectués par des utilisateurs lambda. Instagram n'a fourni un feedback que pour les signalements effectués par



Graphique 3: Taux de feedback utilisateur lambda / trusted flagger par plateforme ; Source : monitoring sCAN

les trusted flaggers. Google+ n'a fourni aucun feedback pendant la période de monitoring. Fournir un feedback aux utilisateurs sur leurs signalements est essentiel pour les impliquer et les motiver à continuer à signaler le contenu illégal.

Expériences et observations

Au cours de la période de monitoring, Facebook, Twitter et Youtube ont reçu le plus grand nombre de signalements de la part des organisations partenaires. C'était la première fois qu'Instagram était inclus dans l'exercice de monitoring et certains partenaires n'avaient qu'une petite expérience dans le monitoring de cette plateforme. Les partenaires ont toutefois rapporté avoir tout de même relevé une quantité importante de discours de haine illégaux sur Instagram, bien qu'il leur ait été plus difficile de trouver du contenu pertinent sur ce réseau. Google+ a fermé en avril 2019 et les partenaires n'ont donc envoyés que quelques signalements. À cause d'une popularité et d'une pertinence se limitant à la Croatie et à la France, Dailymotion et Jeuxvidéo n'ont fait l'objet que d'un petit nombre de signalements.

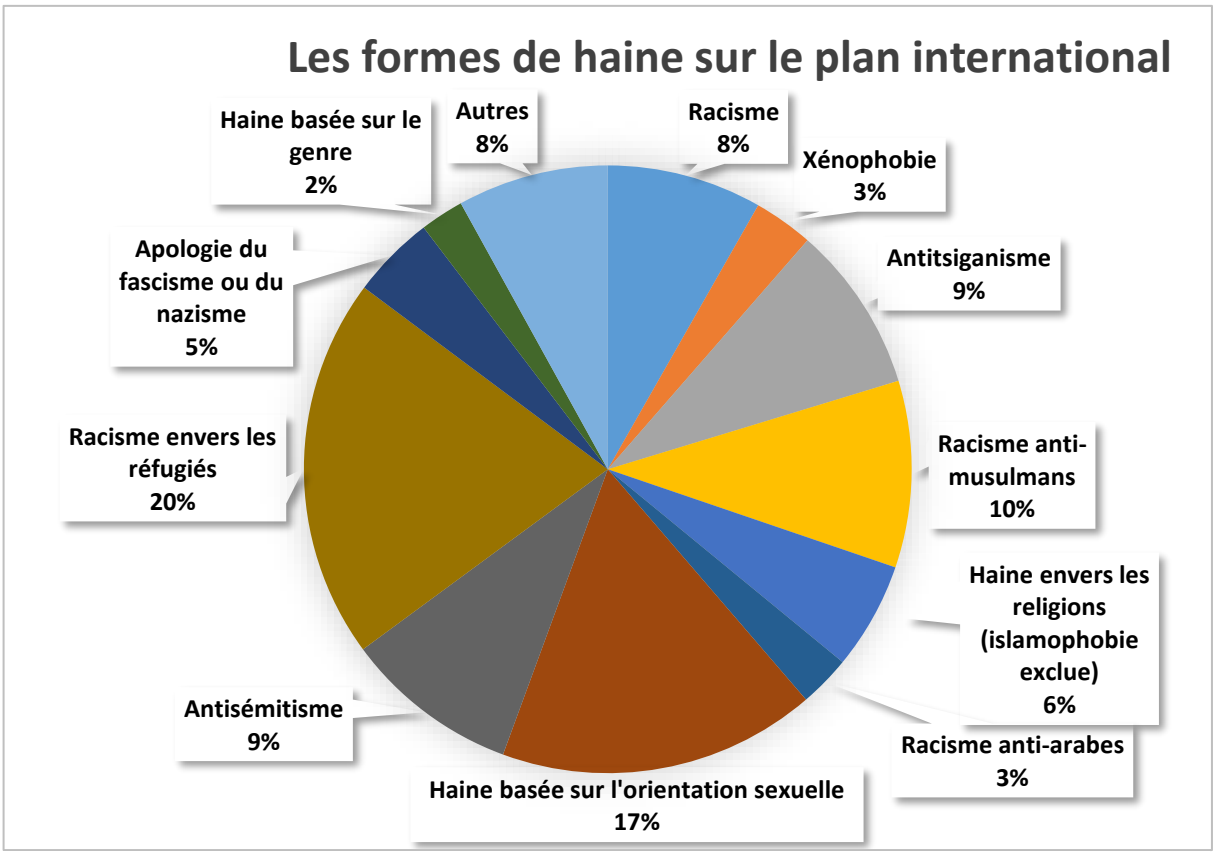
Le bien-être des chercheurs chargés du monitoring était une préoccupation importante pour les partenaires sCAN, qui se sont assurés que leur personnel soit assez formé et soutenu au cours du processus. Afin d'assurer la confidentialité et la sécurité de leurs équipes et de recréer l'expérience d'un utilisateur lambda lors des signalements via les dispositifs publics, les partenaires ont crée des adresses e-mail anonymes et de faux profils sur les plateformes monitorées.

Plusieurs partenaires ont observé que Twitter répondait à certains signalements envoyés en tant que trusted flagger en exigeant davantage d'informations, dont des informations sur la personne effectuant le signalement, sans que les partenaires ne sachent vraiment pourquoi. Ils n'ont pas reçu de feedback supplémentaire ni d'évaluation après avoir fourni les informations demandées, et le contenu est resté en ligne.

Deuxième monitoring : 6 mai – 21 juin 2019

Le deuxième exercice de monitoring a été mis en place du 6 mai au 21 juin 2019. Les partenaires sCAN ont rapporté 432 cas à quatre plateformes de réseaux sociaux différentes, à savoir : Facebook (200 cas), Twitter (107), YouTube (90), et enfin Instagram (35). Les sociétés informatiques ont été alertées via les dispositifs de signalement publics. Après cette première étape, certains partenaires sCAN ont signalé de nouveau 90 cas qui n'avaient pas été retirés, cette fois en tant que trusted flaggers.

Les partenaires ont monitoré un certain nombre de formes de haine au cours de cet exercice, dont certaines plus répandues que d'autres.



Graphique 4: Les formes de haine sur le plan international ; Source : monitoring sCAN

Le diagramme ci-dessus nous donne un aperçu sur les tendances des discours de haine en ligne. D'après l'expérience de nos partenaires, sur la période de six semaines qu'a duré le monitoring, les formes de haine les plus répandues ont été le racisme envers réfugiés, l'homophobie, le racisme anti-musulmans et l'antisémitisme.

Ces résultats sont uniquement représentatifs du contenu relevé par les organisations sur une période précise de six semaines et sur les plateformes surveillées. Certains partenaires se sont concentrés en grande partie sur certaines formes de discours de haine. Les partenaires ont fourni des informations détaillées au sujet des formes de haine qu'ils ont relevé afin de mieux pouvoir évaluer celles-ci.

En règle générale, les formes de haine relevées semblent refléter les tendances générales du discours de haine dans les pays respectifs. Selon l'expérience des organisations partenaires, le discours de haine envers les réfugiés fait partie des plus répandus en Autriche et en Slovénie. En République Tchèque, les Rroms sont la minorité la plus régulièrement attaquée. Depuis 2015, on rencontre plus fréquemment des discours de haine envers les musulmans, les arabes, les réfugiés et les personnes de couleur.

Dans d'autres cas, les discours de haine identifiés au cours du monitoring reflétaient les actuels discours et évolutions dans les pays étudiés. En Italie, il existe clairement une incitation à la haine de la part des autorités. La plupart des contenus haineux signalés par les partenaires italiens était constitué de réactions à des publications ou à du contenu partagé par des politiciens haut placés ou par des partis politiques. La France connaît quant à elle une vague d'antisémitisme depuis le début de l'année. En Croatie, les marches des fiertés qui se sont déroulées pendant la période de monitoring ont surtout été la cible de discours haineux homophobes en ligne.

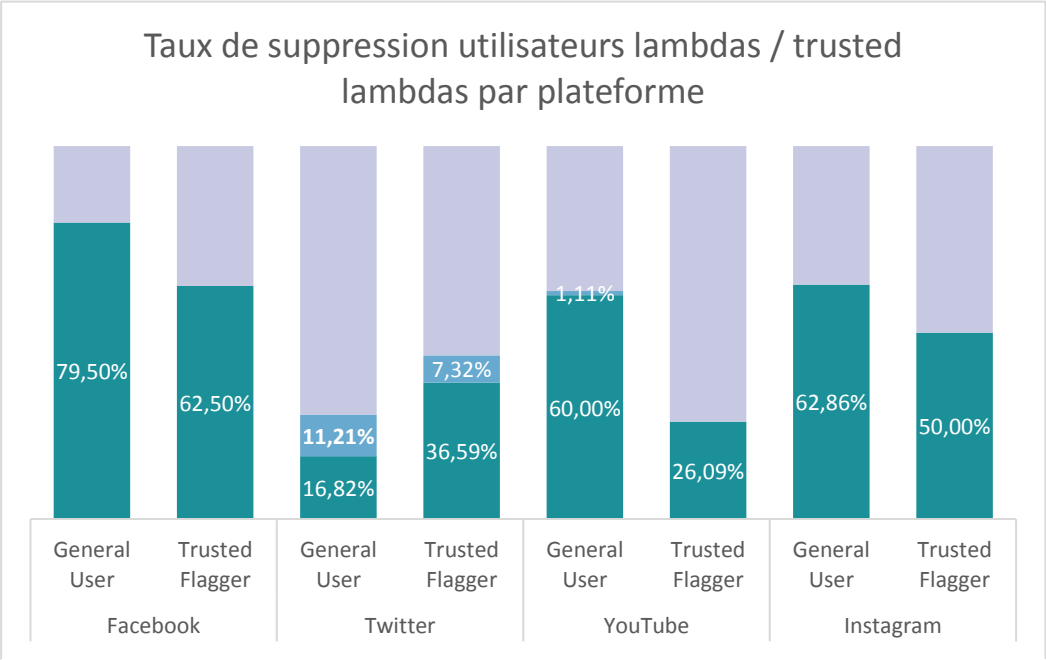
En Lettonie, des discours de haine antisémites ont été déclenchés par un projet de loi sur l'indemnisation de la communauté juive pour perte de propriété communautaire pendant la Shoah, et par le fait que le président letton récemment élu est d'origine lettone et juive. Le rapport sur les crimes de haine envers la communauté LGBT en Lettonie, le projet de loi sur la cohabitation qui a été refusé par le Parlement, ainsi que des attaques envers les personnes homosexuelles en Tchétchénie ont provoqué des discours de haine homophobes. Des discussions au sujet du projet de modification de loi visant à permettre aux étudiants étrangers de travailler à plein temps en Lettonie ont conduit à des discours de haine xénophobes.

Jugendschutz.net, le partenaire allemand, surveille en continu l'extrémisme de droite et l'extrémisme islamiste. Les cas d'islamisme visaient la plupart du temps les personnes qui ne suivaient pas l'idéologie islamiste. La plupart des cas d'extrémisme de droite comportaient des discours de glorification du nazisme. Il convient de noter qu'en Allemagne, l'usage et la dissémination de symboles appartenant à des organisations anticonstitutionnelles est interdit. Les cas allemands qui ont été signalés incluaient donc souvent des symboles associés à des organisations de terrorisme islamiste (comme le drapeau de l'autoproclamé EI) ou des symboles appartenant à des organisations nazies (croix gammées, totenkopf).

Taux de suppression

Au total, les sociétés informatiques ont retiré 67% du contenu signalé au cours du monitoring et ont limité l'accès à 4%. 59% du contenu signalé à travers les dispositifs de signalement disponibles publiquement a été retiré, et 3% a subi des restrictions d'accès. Le reste du contenu n'a été supprimé qu'après avoir été signalé une seconde fois en tant que trusted flaggers. Les sociétés ont réagi dans 43% des cas signalés par les trusted flaggers : 40% de suppression et 3% de restrictions d'accès.

Étonnamment, contrairement aux précédents exercices de monitoring, le taux de suppression n'a pas été supérieur dans le cas des signalements effectués en tant que trusted flaggers, excepté dans le cas de Twitter. Ce dernier a en effet notablement favorisé les signalements des trusted flaggers par rapport à ceux des utilisateurs lambdas.



Graphique 5: Taux de suppression utilisateurs lambdas / trusted flaggers ; Source : monitoring

Néanmoins, il convient de noter que les taux de suppression pour les cas ayant été signalés en tant que trusted flagger étaient assez élevés, bien que ces cas aient déjà été rejetés une fois auparavant, quand ils avaient été signalés en tant qu'utilisateurs lambda. On peut donc déclarer que le signalement en tant que trusted flagger est plus efficace lorsqu'il s'agit des discours de haine que les signalements via les canaux publics disponibles pour tous les utilisateurs sur les plateformes monitorées.

Délais de suppression

Les chiffres varient beaucoup en ce qui concerne les taux et les délais de suppression. Ils dépendent en grande partie du pays et de la plateforme. On peut toutefois dire que de manière générale, Facebook est de loin le plus efficace lorsqu'il s'agit de se débarrasser des contenus haineux et de le faire rapidement et en respectant le Code de conduite.

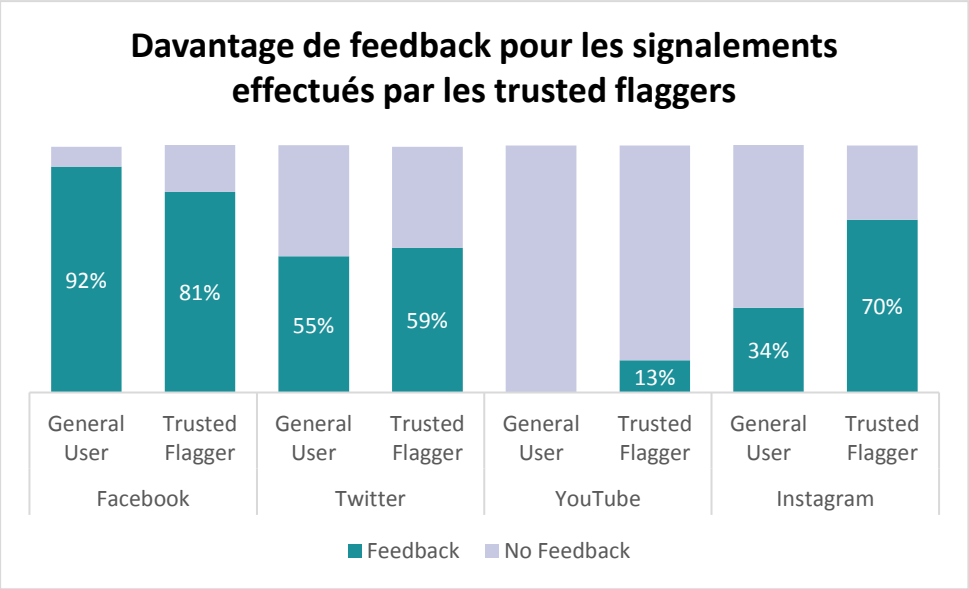
Parmi les cas signalés via les dispositifs de signalement disponibles aux utilisateurs lambdas, Facebook a supprimé 64% en 24 heures. Instagram a retiré ou limité l'accès à 43% des cas en 24 heures, et YouTube 23%.

Twitter n'a retiré ou limité l'accès qu'à 17% du contenu signalé dans ces délais. De plus, il y a des pays où la société n'a supprimé aucun des contenus haineux signalés : Romea en Tchéquie, et UL-FDV en Slovénie ne sont pas parvenus à faire retirer ces contenus de la plateforme. Twitter n'a rien retiré en Slovénie, même lorsque UL-FDV a effectué des signalements en tant que trusted flagger. En Italie, Twitter n'a limité l'accès qu'à un des cas signalés par le CESIE.

La situation est assez similaire pour les trusted flaggers. Facebook est encore une fois de loin la société la plus réactive en termes de suppression de contenu et de suppression en 24 heures (44%). Toutefois, Twitter est bien plus efficace dans ce cas, surtout dans un délai de 24 heures (37%) lorsque le contenu est signalé en tant que trusted flagger. Instagram et YouTube avaient tous les deux un taux de suppression de 30% en 24 heures.

Feedback

Recevoir un feedback sur les signalements effectués sur les réseaux sociaux est extrêmement important à la fois pour les utilisateurs et pour les trusted flaggers. Il est également exigé par le Code de conduite que les sociétés fournissent un réel feedback dans les meilleurs délais. Il y a une énorme différence entre les différentes plateformes en termes de feedback pertinent et rapide.



Graphique 6: Taux de feedback utilisateurs lambdas / trusted flaggers par plateforme ; Source : monitoring sCAN

Facebook est de loin le plus efficace lorsqu'il s'agit de donner une réponse : la société a répondu à 92% des signalements réalisés en tant qu'utilisateurs lambdas (74% en 24 heures). Twitter a répondu à 55% de ces alertes (45% en 24 heures), et Instagram à 34% en 24 heures. YouTube n'a envoyé aucun feedback pour les cas signalés par les partenaires sCAN via les dispositifs de signalement publics.

Comme on peut l'observer sur le graphique ci-dessus, la plupart des plateformes ont fait davantage de feedback aux truster flaggers qu'aux utilisateurs lambda, à l'exception de Facebook qui a répondu plus souvent aux utilisateurs lambda qu'aux trusted flaggers.

Expériences et observations

Afin de mieux appréhender l'expérience des organisations partenaires au cours de la période de monitoring, un questionnaire d'évaluation a été distribué au terme de l'exercice.

Les partenaires ont rapporté que l'accès au contenu n'avait été limité par les sociétés que dans quelques cas, par géo-blocage pour la grande majorité. Toutefois, le partenaire français a rapporté qu'au cours du deuxième monitoring, tous les cas de discours de haine homophobes signalés à Twitter ont été soumis à des restrictions au lieu d'être supprimés. Puisque le contenu reste en ligne et que les méthodes pour contourner les restrictions sont largement connues des utilisateurs, les partenaires sCAN ne considèrent cette approche qu'en partie efficace. Il n'y a eu aucune indication sur les raisons pour lesquelles Twitter n'agissait de la sorte que pour le contenu français homophobe.

Le partenaire tchèque a remarqué que la République Tchèque ne pouvait pas être choisie comme lieu dans la section « tendances » sur Twitter. Ceci complique le processus de monitoring, puisqu'il est alors plus difficile de monitorer les débats nationaux pertinents. Un autre problème rencontré pendant le monitoring était le fait que les liens directs vers les commentaires signalés sur Facebook ne fonctionnaient pas très bien. Dans les longues conversations contenant une multitude de commentaires, il est alors difficile de vérifier si le commentaire a été retiré.

Pour ce qui est du feedback, les sociétés informatiques répondaient principalement par des messages automatiques sans donner de détails à propos du cas en particulier ou du raisonnement justifiant leur

décision. De plus, certaines sociétés informatiques n'ont pas traité les différents partenaires de la même façon. Bien que Facebook ait envoyé le plus de feedback et fournit un feedback à la fois aux utilisateurs lambda et aux trusted flaggers, certains partenaires ont noté que le feedback n'était pas envoyé directement après que la société ait agit, mais seulement quelques jours plus tard.

Le partenaire italien a reçu un feedback personnalisé pour tous les contenus signalés à Facebook, alors que le partenaire slovène n'a reçu que des réponses automatiques lors des signalements effectués en tant qu'utilisateur lambda. Pour les signalements effectués en tant que trusted flaggers, ils recevaient un feedback par mail. Toutefois, ces mails ne référençaient pas toujours le contenu signalé, ce qui rendait le suivi de l'avancement des cas difficile.

Conclusion

En comparant les résultats du premier et du deuxième monitoring, on remarque des taux de suppression similaires. Toutefois, puisque Facebook a reçu le plus de signalements de la part des organisations participantes, ceci ne reflète par l'efficacité de toutes les sociétés informatiques monitorées. En étudiant séparément les réactions des sociétés, on note qu'à l'exception de Facebook, les plateformes n'ont pas été aussi efficaces lors du deuxième monitoring. En ce qui concerne le contenu signalé en tant que trusted flaggers, moins de cas ont été supprimés ou ont subi des restrictions d'accès par rapport aux précédents exercices de monitoring organisés par la Commission européenne.

Dans le Code de conduite, les entreprises s'engagent à examiner et à retirer le contenu qui va à l'encontre de la loi nationale ou de leurs conditions générales sous 24 heures. Toutefois, seuls Facebook et Instagram ont réussi à atteindre un niveau acceptable de suppression de contenu haineux signalé dans ces délais, et Twitter et YouTube n'ont pas atteint les 50%.

Les sociétés se sont également engagées à fournir un feedback pertinent dans les meilleurs délais aux personnes signalant du contenu illégal. La suppression des contenus haineux est importante, mais le feedback l'est tout autant si ce n'est davantage. Fournir un feedback, même s'il s'agit d'une réponse automatique, est considéré comme crucial dans une optique de transparence envers les utilisateurs et cela les encourage également à soutenir les réseaux sociaux dans la lutte contre les discours de haine en ligne.

Les partenaires ont observé une baisse des feedbacks par rapport aux précédents exercices de monitoring. Facebook est la seule société à avoir systématiquement fourni un feedback à la fois aux utilisateurs lambda et aux trusted flaggers au cours des deux exercices de monitoring. Facebook est également la seule société à avoir envoyé la majorité des feedbacks dans les 24 heures suivant le signalement. YouTube a été particulièrement critiqué, ne fournissant en effet presque aucun feedback, ni aux utilisateurs lambda ni aux trusted flaggers.

Afin d'appliquer pleinement le Code de conduite et de combattre efficacement les discours de haine en ligne, il est fondamental que les plateformes de réseaux sociaux réagissent à tous les signalements qu'ils reçoivent de la part de leurs utilisateurs dans les meilleurs délais, peu importe qui signale et quand. La poursuite des actions de monitoring et de lutte sont essentielles pour assurer un espace en ligne sûr et respectueux en Europe et ailleurs.