Rapport Annuel

Juliet 2019 – Avril 2020

Platforms, Experts, Tools:
Specialised Cyber-Activists Network

Rapport Annuel

Juliet 2019 – Avril 2020

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 ».

Clause de non-responsabilité

Sommaire

À propos du projet **3**

Introduction **6**

Outils et ressources **8**

L'exercice d'expérimentation de sCAN *10*

L'intégration de l'intelligence artificielle *10*

Guide pour l'utilisation des technologies automatisées dans la surveillance des contenus de discours de haine *12*

Recherches **13**

Analyse « Hotspot de la haine » *14*

Analyse « Les discours de haine intersectionnel en ligne » *15*

Analyse « Discours de haine et la pandémie aux temps d'Internet » *16*

Les exercices de monitoring de sCAN **18**

Éducation **24**

Formation en ligne sur les discours de haine *25*

Cours de modération en ligne *26*

Formation avancée en monitoring *26*

Perspectives futures et recommandations politiques **28**

Ressources et lectures complémentaires **31**

Le ressources du projet sCAN *31*

Bibliographie *32*

Introduction

Internet fait partie intégrante de la communication quotidienne mondiale et, bien qu'il soit principalement utilisé pour communiquer entre amis ou pour partager des opinions sur des sujets variés, certains utilisateurs s'en servent toutefois pour disséminer la haine et inciter à la violence à l'encontre des minorités vulnérables. La nature globale d'Internet et son interconnexion mondiale vont dans le sens d'une approche internationale de la lutte contre les discours de haine en ligne.

Ces dernières années, plusieurs projets européens de lutte contre les discours de haine ont été mis en œuvre. Afin d'encourager la mise en réseau à l'échelle européenne et de tirer parti de la complémentarité des résultats de différents projets, le projet sCAN entretient une étroite collaboration avec l'INACH (International Network Against Cyber Hate) et le projet Facing Facts!

Les partenaires du projet s'entendent sur la définition du discours de haine proposée par l'INACH :

> « Le discours de haine comprend les propos publics de nature volontairement ou involontairement discriminatoire et/ou diffamatoire; l'incitation intentionnelle à la haine et/ou à la violence et/ou à la ségrégation fondée(s) sur la race, l'ethnicité, la langue, la nationalité, la couleur de peau, les croyances religieuses ou leur absence, le genre, l'identité de genre, le sexe, l'orientation sexuelle, les opinions politiques, le statut social, la propriété, la naissance, l'âge, la santé mentale, le handicap ou la maladie d'une personne ou d'un groupe de personnes, que ces caractéristiques soient perçues ou réelles. » [1]

Pendant la deuxième année de mise en œuvre du projet (juin 2019 - avril 2020), les partenaires ont poursuivi leur approche multidimensionnelle de recherche de solutions technologiques, de recherche, de monitoring et d'éducation. Le projet a mené des exercices de monitoring pour évaluer

[1] International Network Against Cyber Hate (2018). *What is cyber hate*. Disponible sur **http://www.inach.net/wp-content/uploads/WHAT-IS-CYBER-HATE-update.pdf** (consulté le 24.06.2019).

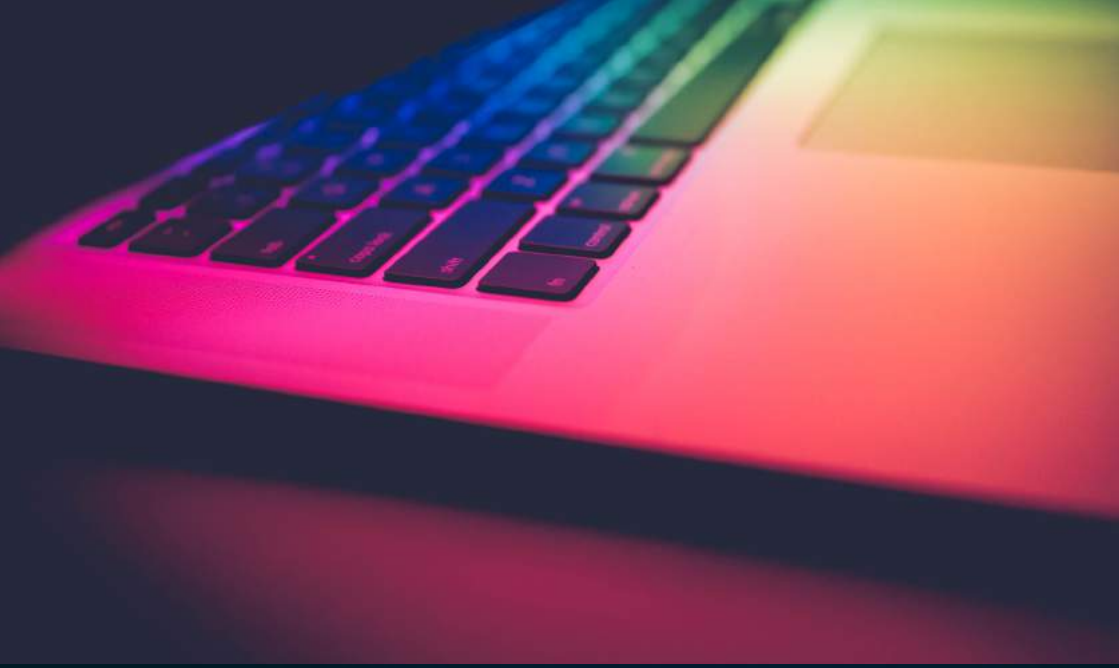
l'efficacité de certains crawlers en ligne et de l'intelligence artificielle pour faciliter la surveillance des discours de haine en ligne.

En outre, les partenaires ont mené des projets de recherche communs et publié des rapports analytiques sur leurs résultats. Tout d'abord, les partenaires ont analysé les discours de haine diffusés ou facilités par des personnalités publiques, telles que des journalistes et des personnes influentes dans la politique et enfin dans les réseaux sociaux. Dans le cadre du deuxième projet de recherche, les partenaires ont analysé les discours de haine intersectionnels en ligne dans les différents pays participant au projet. Un troisième projet de recherche analysera l'impact de la pandémie actuelle de Covid-19 sur les discours de haine en ligne.

En outre, les partenaires du projet 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 illégaux sur Internet, élaboré en 2016 par la Commission européenne. Les partenaires de sCAN ont déjà participé à des exercices de monitoring antérieurs organisés par la Commission européenne et l'INACH.

Toutefois, la recherche et la surveillance ne suffisent pas à elles seules pour lutter contre la cyberhaine. C'est pourquoi les partenaires ont développé et conduit des cours en ligne et des ateliers de formation hors ligne pour renforcer les capacités des organisations de la société civile (OSC) et des militants individuels à contrer les discours de haine de diverses manières, par la surveillance, le contre discours ou la modération des discussions en ligne.

Tous les résultats du projet sont accessibles sur le site www.scan-project.eu. Des liens directs aux documents seront fournis à la fin de ce report.



Outils et ressources

L'un des principaux objectifs du projet sCAN est d'aider et de fournir de nouveaux outils de surveillance afin de faciliter les efforts de recherche pour lutter contre les discours de haine en ligne.

Cette détection pourrait être assurée par plusieurs types d'outils tels que le *web spidering, les crawlers,* les logiciels et l'intelligence artificielle. Toutefois, la plupart de ces outils ne sont pas facilement accessibles aux OSC et aux cyberactivistes. En effet, certaines conditions doivent être prises en compte par les organisations qui souhaitent utiliser des outils de surveillance automatisés: ressources humaines, matérielles, ressources et défis linguistiques et existence d'un bureau informatique ou de compétences de codage. En plus de ces éléments, il est crucial de comprendre que toutes les plateformes en ligne ne fournissent pas les mêmes options concernant la détection des discours de haine: leur niveau de confidentialité a un impact direct sur les possibilités d'utilisation des technologies automatisées. Une autre observation importante est que les outils de

surveillance automatisés ne doivent pas être considérés comme le seul moyen efficace de lutter contre les discours de haine en ligne. L'expertise humaine en termes de connaissances, de capacité d'adaptation et de compétences d'analyse reste cruciale pour la surveil-lance des discours de haine.

Au cours de la première année de ce projet, les partenaires de sCAN ont contribué à la constitution d'un ensemble de mots et d'expressions clés dans toutes les langues du projet, y compris des informations supplémentaires sur le contexte dans lequel ces mots sont utilisés dans les discours nationaux respectifs. Cette recherche a fourni des informations importantes sur la nature des discours de haine dans les pays analysés. Pour compléter ces résultats, une étude de cartographie a été réalisée, identifiant certaines des solutions logicielles et des outils disponibles pour surveiller automatiquement la cyberhaine. Pour les OSC, il est important d'explorer l'utilisation d'outils logiciels automatisés pour surveiller les discours de haine. Néanmoins, certaines conditions doivent être remplies : les *crawlers*, les logiciels ou l'intelligence artificielle doivent être fournis avec un ensemble spécifique de mots-clés tenant compte du contexte national et des modèles de discours de haine dans chaque pays. D'autres critères sont également essentiels pour garantir l'intégration d'outils automatisés dans la surveillance des discours de haine par les OSC. Il s'agit notamment des coûts d'utilisation d'un certain outil ou des compétences techniques requises pour le faire fonctionner.

Au cours de la deuxième année du projet, l'expérimentation lancé en septembre 2018 a été poursuivi. En outre, le projet a développé un partenariat avec Factmata, une société spécialisée dans l'intelligence artificielle. Deux sessions de campagnes de tests ont été programmées tout au long du projet : la première dédiée aux crawlers et la seconde à l'intelligence artificielle.

L'exercice d'expérimentation de sCAN

Au cours du projet sCAN, jugendschutz.net et Licra ont développé une méthodologie commune pour tester une sélection d'outils automatisés. L'objectif principal était de fournir une évaluation de l'efficacité et de la pertinence des outils sélectionnés afin de les intégrer dans la tâche de monitoring du consortium SCAN. Cette expérimentation a été organisé dans le cadre de deux campagnes différentes.

La première campagne de test a été consacrée aux crawlers au début du projet, en septembre et octobre 2018. Cette campagne de deux mois était axée sur le test de plusieurs crawlers sur des sites web, des blogs et des plateformes de réseaux sociaux en utilisant des mots clés pertinents sélectionnés pour le rapport sur les ontologies de la haine. Les partenaires se sont concentrés sur les plateformes de réseaux sociaux qui ont signé le code de conduite sur la lutte contre les discours de haine illégaux en ligne avec l'UE et ont également été inclus dans les exercices de monitoring du projet.

Au cours de cette première campagne, les outils suivants ont été testés : TAGS v 6.1 sur Twitter, HTTracks pour le web 1.0, SociScraper sur Instagram et YouTube et CrowdTangle sur Facebook. Chaque outil sélectionné a été testé selon une méthodologie commune incluant l'utilisation d'une liste de mots-clés sélectionnés sur la base de l'ontologie de la haine publiée au cours de la première année du projet. En outre, une liste non exhaustive de critères a été établie pour évaluer les outils sélectionnés : le prix du crawler/logiciel, la formation, les compétences requises, l'assistance manuelle, les résultats concernant les catégories de discours haineux, le temps, les bugs et les problèmes, les avantages et les inconvénients, les paramètres linguistiques.

L'intégration de l'intelligence artificielle

L'intelligence artificielle (IA) s'impose dans la détection des discours de haine en ligne. **Pour cette raison** la deuxième campagne de tests s'est concentrée sur les algorithmes et l'intelligence artificielle à partir de décembre 2018

jusqu'à la fin de la mise en œuvre du projet sCAN. Comme le développement de l'AI nécessite des ressources difficilement accessibles pour les OSC et les cyber-activistes des droits de l'homme, un partenariat a été formé avec la société Factmata.

Factmata travaille sur une technologie combinant intelligence artificielle, algorithme et connaissances d'experts pour faire face aux discours de haine et aux fausses nouvelles. Basée à Londres, Factmata propose une plateforme d'intelligence artificielle (API) et des services de lutte contre les fausses nouvelles en fournissant un système de notation du contenu sur le web. En ce qui concerne le contenu des discours de haine, l'API de Factmata note le contenu en fonction de critères : « insulte », « obscénité », « toxicité », « stéréotype », et « menace », « haine de l'identité » ainsi que « sexisme » et contre « tout genre particulier ». Pour affiner leur algorithme, la start-up a besoin d'un soutien humain afin d'améliorer la détection des discours de haine et les résultats. jugendschutz.net et Licra ont contribué, en tant que membres de la communauté des utilisateurs, à vérifier la qualité des contenus à l'aide de leurs outils d'intelligence artificielle.

jugendschutz.net, Licra et Factmata ont lancé leur partenariat en novembre 2018. Pendant plusieurs sessions de test de l'IA, jugendschutz.net et Licra ont participé à des sessions d'annotation concernant les critères de discours de haine, de menaces, d'insultes et d'obscénité. L'outil a été formé dans un contexte nord-américain. En conséquence, les deux organisations européennes ont fourni une expérience locale sur la façon de définir le discours de haine : elles ont contribué à intégrer les critères transnationaux européens et nationaux des tendances en matière de discours de haine et, par conséquent, à améliorer le modèle de détection des contenus haineux. En raison de l'évolution très rapide du vocabulaire des discours de haine, cette méthodologie s'efforce d'aborder les questions de langage et d'évolution. Pour la deuxième campagne, le test a été réalisé avec un contenu en anglais uniquement, car l'algorithme n'a pas été formé sur d'autres langues. Néanmoins, grâce au partenariat développé avec Factmata, il a été possible de former l'outil dans d'autres langues.

Guide pour l'utilisation des technologies automatisées dans la surveillance des contenus de discours de haine

Les résultats de ces campagnes ont été analysés afin de produire un guide d'utilisation complet sur les outils de surveillance automatisés. Dans ce guide, le consortium sCAN vise à expliquer comment utiliser les outils disponibles pour améliorer la surveillance et le retrait des discours haineux.

Le guide de l'utilisateur fournit des lignes directrices sur les outils disponibles et peu coûteux avec une interface complète afin d'améliorer le processus de surveillance et la collecte de données. Pour chaque outil, les lignes directrices sont présentées selon le même schéma: présentation; conditions d'utilisation; étapes illustrées pour l'utilisation; et avantages et inconvénients de l'outil.

En outre, deux webinaires sur le contenu du guide ont été organisés pour le consortium afin de faciliter l'utilisation et l'intégration des outils automatisés sélectionnés. De plus, un tutoriel interne en ligne sur l'utilisation de la base de données INACH lors du deuxième exercice de monitoring a été organisé en mai 2019.

Recherches

Tout au long du projet sCAN, les partenaires se sont consacrés à la rédaction de documents analytiques en mettant l'accent sur des sujets actuels et controversés de la plus haute importance. Ces sujets ont été choisis sur la base de l'expertise et des expériences du consortium et ont été produits dans le but de partager les connaissances et de fournir une vue d'ensemble des tendances et des développements importants du phénomène du discours de haine en ligne. Entre juillet 2019 et avril 2020, le consortium sCAN a publié deux documents analytiques.

Étant donné que la portée du projet sCAN ne permettait pas la mise en œuvre d'analyses qualitatives et/ou quantitatives approfondies, le consortium a décidé de se concentrer sur des études de cas exemplaires afin de fournir une vue d'ensemble et une compréhension approfondie des phénomènes en discussion.

Analyse

« Hotspot de la haine »

Le troisième des quatre documents analytiques du projet était consacré au thème « Les Hotspot de la haine - la responsabilité en ligne des personnalités publiques ». Ce document était basé sur l'expérience des partenaires du projet qui ont pris conscience que les personnalités publiques telles que les politiciens et les journalistes peuvent avoir une forte influence sur leurs adeptes sur les réseaux sociaux et dans la sphère en ligne. En examinant de plus près les études de cas, le consortium a constaté que ces « influenceurs » communiquent souvent en permanence avec leurs partisans via les réseaux sociaux et ont donc la possibilité de façonner considérablement leurs perceptions. Avec leur portée énorme et la quantité de contenu, les réseaux sociaux sont l'outil parfait pour influencer l'opinion publique. Leur grande portée est la raison pour laquelle les réseaux sociaux portent une responsabilité particulière lorsqu'il s'agit de diffuser des désinformations ou des incitations implicites (ou parfois explicites) à la haine. Dans plusieurs pays européens, des personnalités politiques et publiques de premier plan utilisent leur présence en ligne pour inciter à la haine ou pour encourager les discours haineux en publiant des commentaires tendancieux et populistes sur leur profil dans les réseaux sociaux. Même si les messages eux-mêmes ne constituent pas un discours de haine illégal, ils incitent à la haine et stimulent le discours de haine dans les sections de commentaires. Il peut être difficile de contrer de tels cas de déclenchement de discours de haine tout en respectant la liberté d'expression. Si le message original reste en ligne, il est susceptible d'attirer d'autres commentaires haineux.

L'analyse des études de cas de tous les pays participant au projet a montré clairement que les sociétés de réseaux sociaux doivent examiner ces cas de près et commencer à explorer les moyens de les traiter à grande échelle. L'incitation fonctionne comme une simple allumette qui enflamme toute une forêt. Les hotspot de la haine en ligne doivent donc faire l'objet d'une attention particulière de la part des réseaux sociaux : ils ne peuvent être ni ignorés, ni sous-estimés et même les hommes politiques et les organes d'information doivent prendre plus au sérieux leur responsabilité de modérer les commentaires sur leurs profils et leurs chaînes. En outre, les utilisateurs peuvent contester les expressions haineuses par des contre-discours, en déconstruisant les

stéréotypes haineux et en démystifiant les fausses nouvelles et les manipulations. Ils peuvent également choisir d'exprimer leur solidarité et leur soutien aux personnes et communautés ciblées.

Analyse

« Les discours de haine intersectionnel en ligne »

Le quatrième document a comme sujet « Le discours de haine intersectionnel en ligne ». Le concept de discrimination intersectionnelle trouve son origine dans le mouvement du féminisme noir et le terme « intersectionnalité » a été inventé par Kimberlé Crenshaw. Toutes les organisations impliquées dans le projet sCAN perçoivent le phénomène de l'intersectionnalité comme un défi constant dans la sphère analogique et en ligne. Les organisations de sCAN ont examiné de plus près l'intersectionnalité et ont suivi la suggestion de la Fundacion Secretariado Gitano d'analyser comment un incident de discrimination spécifique aurait été différent si l'une des caractéristiques d'intersection avait été absente.

L'analyse d'un certain nombre de cas a permis d'établir que le discours haineux intersectionnel est courant dans tous les pays participant au projet (et probablement au-delà). Dans l'ensemble, les femmes perçues*, les personnes LGBTIQ+ perçues et/ou les personnes affiliées ou appartenant à une minorité ethnique et/ou religieuse - sur la base d'une combinaison de leurs catégories d'identité (légalement protégées) ont été identifiées comme les groupes cibles les plus fréquents. En outre, les personnes présentant des caractéristiques visibles ainsi que celles occupant des postes publics se sont avérées particulièrement touchées par le discours haineux intersectionnel.

Les départements gouvernementaux devraient consacrer le principe de l'intersectionnalité dans toutes les politiques d'égalité, afin de ne pas négliger les expériences des groupes les plus touchés par la discrimination intersectionnelle. Les gouvernements devraient mettre en place des mécanismes de consultation solides avec un large éventail de parties prenantes diverses. Les politiciens et les hauts fonctionnaires des autorités publiques devraient condamner fermement le discours de haine et promouvoir le contre-discours en mettant l'accent sur la discrimination multiple et le phénomène de la haine intersectionnelle en ligne et son impact sur les personnes directement

concernées. En outre, tous les partis politiques devraient condamner les discours discriminatoires, incendiaires ou haineux, en mettant l'accent sur la discrimination multiple, et appeler leurs membres et leurs partisans à s'abstenir d'utiliser des discours haineux pendant les campagnes électorales.

L'utilisation de discours haineux pour créer une atmosphère d'intolérance et d'exclusion pour un groupe de notre société peut déclencher des incidents violents. Par conséquent, les autorités chargées de l'application de la loi devraient non seulement garantir une enquête adéquate sur les discours de haine et autres incidents discriminatoires, mais aussi être conscientes et prendre en compte des facteurs aggravants tels que, par exemple, la couleur de la peau, l'orientation sexuelle, l'identité sexuelle, l'identité de genre, les handicaps, l'âge et la religion. Les autorités répressives devraient renforcer leur coopération avec divers groupes et communautés (socialement construits) afin de mieux comprendre comment certains groupes et communautés sont affectés par le discours de haine, en particulier le discours de haine intersectionnel.

Le discours de haine intersectionnel est encore plus difficile à classer et à combattre que le discours de haine ciblant une caractéristique réelle ou perçue. Nous avons tendance à utiliser des outils de lutte contre la haine conçus pour une forme spécifique de discours de haine et, lorsque plusieurs formes de discours de haine se croisent, certaines expressions peuvent être sous-rapportées, simplifiées ou même ignorées. Les OSC devraient donc renforcer leurs efforts pour signaler et contrer les discours de haine intersectionnels.

Analyse

« Discours de haine et la pandémie aux temps d'Internet »

Une analyse supplémentaire a été consacré au thème « Discours de haine et la pandémie aux temps d'Internet ». La pandémie de Covid-19 a perturbé les conditions de vie sociale, économique et politique des personnes dans toute l'UE. Alors que la pandémie se développe, les phénomènes en ligne de théories du complot, de rumeurs, de fausses nouvelles et de contenus haineux liés à cette maladie mondiale se multiplient. En raison des procédures de confinement, les gens passeront probablement plus de temps en ligne,

et discuteront et interagiront par l'intermédiaire des réseaux sociaux.

Les discours haineux en ligne que l'on trouve dans la pandémie suivent des schémas traditionnels mais peutêtre même nouveaux. Les mécanismes de désignation de boucs émissaires et la propagation de rumeurs entraînent une large diffusion. Ce phénomène social et psychologique bien connu a déjà été observé au cours d'épisodes de pandémie mondiale dangereux antérieurs, comme par exemple pour la peste noire au Moyen-Âge. Presque chaque crise sociale, économique et sanitaire peut entraîner la montée des théories du complot, y compris des croyances haineuses. Les complots découverts donnent accès à la « rationalité » et aux « phénomènes explicables ». Des crises comme les pandémies peuvent causer ou approfondir une division dans les sociétés par la propagation de ragots, de théories du complot, d'accusations et, par conséquent, d'actes violents contre « l'Autre ».

Pour toutes ces raisons, le consortium a partagé ses expériences, ses connaissances et quelques explications clés sur les mécanismes possibles de l'interaction entre les discours de haine et la montée d'une pandémie mondiale à l'ère d'Internet et des réseaux sociaux. Ce document analytique vise à analyser les tendances haineuses en ligne en période de pandémie, y compris dans une perspective historique. Les principaux objectifs de ce rapport sont d'identifier certains des événements qui ont entraîné une interaction entre une pandémie et une montée des discours et des actes haineux, afin de mieux expliquer et de s'attaquer aux stéréotypes et aux théories d'aujourd'hui concernant cette nouvelle crise sanitaire à laquelle notre monde est confronté.

Les exercices de monitoring de sCAN

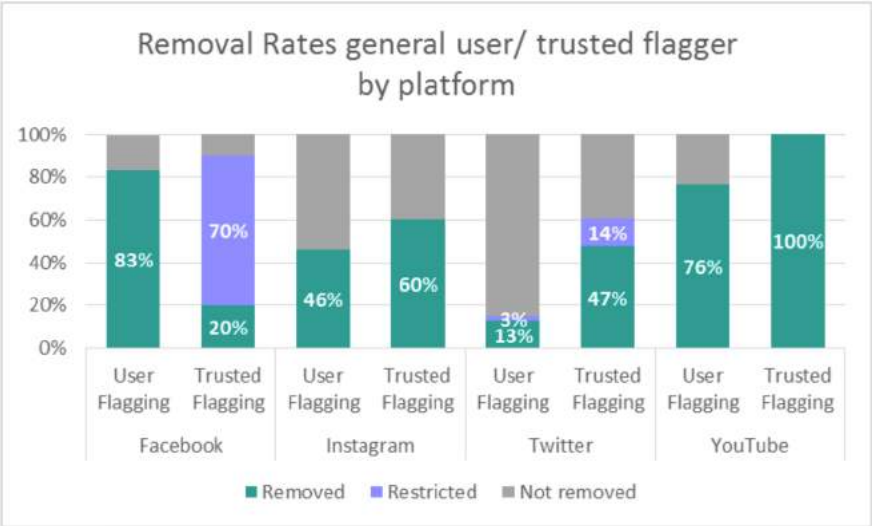
Au cours de la deuxième année de mise en œuvre du projet, les organisations partenaires de sCAN ont participé à deux exercices de monitoring, l'un avec la Commission européenne et l'autre avec le Réseau international contre la haine cybernétique (INACH) et le projet Open Code for Hate-Free Communication (OpCode). Le but de ces exercices était d'évaluer la conformité de Facebook, Twitter, YouTube et Instagram avec le code de conduite de l'UE sur la lutte contre les discours haineux illégaux en ligne. Les partenaires de sCAN ont déjà participé à des exercices de monitoring antérieurs organisés par la Commission européenne et l'INACH.

Le troisième monitoring global de sCAN a été mené pendant l'exercice de monitoring organisé par la Commission européenne du 4 novembre au 13 décembre 2019. Au cours de cette période de six semaines, les parte-

naires de sCAN 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 des partenaires de sCAN (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 n'avoir pas été retirés dans la semaine suivant le rapport initial par les canaux de signalement des utilisateurs généraux. Twitter a reçu 59 rapports de signalement, Facebook et Instagram ont reçu chacun 10 rapports et YouTube a reçu 5 par le biais de canaux de signalement. Aucun cas n'a été transmis à Dailymotion et à Jeuxvideo.

Dans l'ensemble, 67,56 % du contenu n'était plus disponible à la fin du monitoring dans le pays d'où il avait été signalé (64,25 % ont été supprimés, 3,31 % ont été restreints). 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



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

première notification par les canaux normaux des utilisateurs (57,80 % ont été supprimés, 0,94 % ont été restreints). Certains partenaires ont fait remonter le contenu qui n'avait pas été supprimé dans la semaine suivant 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 (48,81 % ont été retirés, 17,86 % ont été restreints).

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 dans les 24 heures.

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

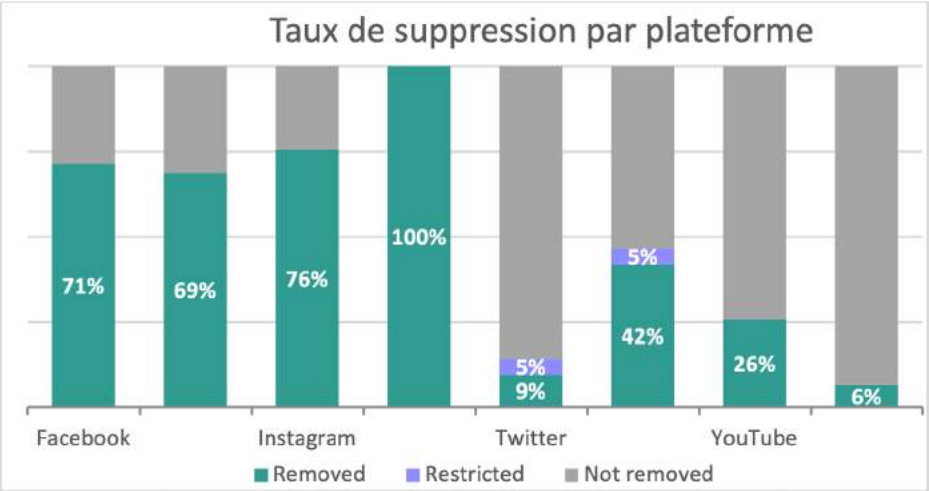
Toutes les plateformes ont obtenu de bien meilleurs résultats pour les rapports soumis par des canaux de signalement fiables. YouTube a supprimé 100 % des rapports soumis par des trusted flaggers. Facebook a pris des mesures dans 90 % des cas en limitant 70 % et en supprimant 20 % des signalements. Les partenaires du projet ne comprennent pas bien pourquoi ils ont choisi de limiter 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 observée pour Twitter. L'entreprise a pris des mesures dans 61 % des cas (47 % ont été supprimés, 14 % ont été restreints), ce qui est presque quatre fois plus que les mesures prises dans les cas signalés par les canaux disponibles pour les utilisateurs généraux.

Le quatrième monitoring de sCAN a eu lieu entre le 20 janvier 2020 et le 28 février 2020. Il s'agissait d'un contrôle « inopiné » en coopération avec le secrétariat de l'INACH et le projet OpCode. Les partenaires de sCAN 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 supprimés après la notification en

tant qu'utilisateurs généraux ont été signalés à nouveau par des canaux de notification disponibles uniquement pour les trusted flaggers.

Dans l'ensemble, seuls 58 % des cas signalés n'étaient plus disponibles à la fin de la surveillance. Il s'agit là d'une baisse importante par rapport au troisième exercice de monitoring de sCAN 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.



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

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 retirés par les sociétés informatiques. Instagram a supprimé tous les cas qui lui avaient été signalés une deuxième fois par les canaux de signalement fiables. 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 trusted flaggers (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.

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.

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

La plupart des entreprises fournissent davantage de commentaires aux signataires de confiance qu'à leur base d'utilisateurs générale. Cela peut être problématique, car les OSC reconnues comme des trusted flaggers ne peuvent pas surveiller et signaler tous les discours de haine illégaux par elles-mêmes. Dans le cas d'Instagram, le dispositif utilisé pour le signalement semble également avoir une incidence sur la réception ou non d'un retour d'information. Alors que les partenaires faisant des rapports par l'intermédiaire de l'application mobile ont

déclaré avoir reçu un retour d'information de la plateforme, les partenaires faisant des rapports à Instagram en utilisant un ordinateur de bureau n'ont pratiquement pas reçu de retour d'information.

L'implication de tous les utilisateurs des plateformes dans le signalement des discours de haine est cruciale 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 leurs normes communautaires.

Éducation

Comme la monitoring et la recherche ne suffisent pas à elles seules pour lutter contre les discours haineux en ligne, le projet sCAN a mis au point des cours de formation en ligne et hors ligne. Ces cours, déjà élaborés et mis en œuvre au cours de la première année du projet, ont également été menés et constamment affinés au cours de la deuxième année.

Les cours en ligne se sont concentrés sur la fourniture de connaissances générales sur les discours de haine, la législation nationale, européenne et internationale en la matière, le contrôle du contenu des discours de haine en ligne, le discours national et la modération des discussions en ligne.

En outre, deux formations avancées sur la surveillance hors ligne ont été organisées à Vienne et à Bruxelles. Elles comprenaient des sessions interactives sur la manière de reconnaître les discours de haine, l'importance de la surveillance et l'art de la documentation.

Formation en ligne sur les discours de haine

L'une des activités du projet sCAN a permis de réaliser le projet Facing Facts! sur les discours de haine en ligne en allemand et en français et de l'adapter aux contextes nationaux respectifs. Ces cours ont été développé pour toute personne intéressée par la lutte contre le discours de haine en ligne en fonction de ses possibilités et capacités. Ils offrent de nouvelles perspectives et des approches pratiques pour lutter efficacement contre les discours de haine en ligne pour un large éventail de personnes telles que les activistes individuels, les membres de communautés, les représentants d'OSC ou les autorités.

Ils fonctionnent avec des méthodes interactives et donnent des informations sur le concept de discours de haine et sur la manière de reconnaître sa nature et ses effets. On peut apprendre à surveiller les discours de haine sur l'internet et comment la surveillance peut être un outil pour contrer le phénomène. Les contre-discours, les contre-campagnes et les contre-récits actifs sont également abordés et le cours montre lesquelles de ces stratégies sont les plus adaptées aux objectifs spécifiques d'un « contre-activiste ».

Les cours en langue allemand sur les discours de haine, intitulé « *Hate Speech* - was tun ? », aborde le contexte allemand et autrichien des discours de haine en ligne. Entre juillet 2019 et avril 2020, le cours en ligne a été proposé à trois reprises. Chaque cours a été proposé à un groupe stable de participants qui se sont impliqués pendant une période de six à huit semaines. Cette approche visait à activer les participants au cours et leur a donné l'occasion d'interagir intensivement entre eux et avec les deux tuteurs en ligne du cours. L'un des tuteurs était du partenaire autrichien ZARA et l'autre du partenaire allemand jugendschutz.net. Ils ont animé le forum de discussion en ligne et ont donné des informations sur les détails et des réponses aux questions qui se posaient. Les cours ont été complétés par des webinaires en ligne avec des experts invités de « Gegen Vergessen - Für Demokratie e.V. » et des comités « No Hate Speech » d'Autriche et d'Allemagne. Au total, plus de 200 personnes ont participé à la version allemande de la campagne « Facing Facts ! Cours en ligne « *Hate speech*- Was tun ? »

Cours de modération en ligne

S'appuyant sur les cours général sur les discours de haine, le projet sCAN a développé un cours en ligne sur la modération des discours de haine en ligne. Ces cours sont disponibles en anglais et en français sur la plateforme Facing Facts Online ! Ils s'adressent aux activistes, aux leaders de communautés, aux blogueurs, aux vlogueurs et aux praticiens qui souhaitent encourager les échanges pacifiques en ligne, mais tout acteur intéressé par le sujet peut les suivre.

Grâce à des outils interactifs, des vidéos, des conférences dynamiques, des études de cas, des témoignages et des quiz, elle aborde la question de la réaction aux commentaires haineux dans les discussions en ligne. Il vise à faire mieux comprendre les principes directeurs de la modération en ligne et les outils qui la soutiennent.

Afin de maintenir des conversations saines en ligne, les cours abordent les différentes options d'intervention, de la suppression au contre-discours, et il encourage également les participants à créer leurs propres politiques de modération basées sur les valeurs qu'ils apprennent à articuler pendant ce parcours.

Formation avancée en monitoring

Une formation hors ligne sur le monitoring avancée et la lutte contre la haine en ligne a été élaborée et mise en œuvre pendant toute la durée du projet. Les participants ont eu la possibilité de devenir des experts dans le domaine de la surveillance et de la lutte contre les discours haineux, de documenter le phénomène, de s'attaquer à la sous-déclaration, de comparer les résultats en ce qui concerne les données acquises au cours des différents exercices et phases de monitoring, ainsi que d'appliquer un système efficace de rapports sur les droits de l'homme.

Les formations comprenaient des sessions interactives sur la manière de reconnaître le discours de haine, l'importance du suivi et l'art de la documentation. Un formateur expert du Réseau international contre la cyberhaine (INACH) a en outre proposé des sessions matinales d'une heure sur la manière d'utiliser la base de données INACH pour documenter la cyberhaine. Les cours s'adressaient aux militants et aux organisations de la société civile qui prévoient de lancer leur propre surveillance des discours de haine en ligne ou qui cherchent à professionnaliser les efforts de surveillance déjà existants.

Les deux formations sur la surveillance avancée et la lutte contre la haine en ligne, mises en œuvre tout au long de la deuxième année du projet (juillet 2019 à avril 2020), ont eu lieu à Vienne (octobre 2019) et à Bruxelles (mars 2020). Elles ont été organisées par les experts en formation de ZARA et la formation à Bruxelles a été co-organisée par CEJI - A Jewish Contribution to an Inclusive Europe. 34 participants de 10 pays européens différents ont eu l'occasion de réfléchir au phénomène de la haine en ligne au sein de groupes transnationaux, d'acquérir des connaissances et de l'expertise, de rassembler des exemples de bonnes pratiques et de créer des alliances et des réseaux solides afin de lutter contre la haine en ligne. 70 personnes ont participé aux quatre formations mises en œuvre tout au long de la période du projet.

De plus, les participants ont eu la possibilité de s'impliquer dans des activités d'auto-sensibilisation pour comprendre et différencier les différentes formes de haine, de discrimination et de cyber-mobbing en ligne.

Afin de garantir la durabilité des connaissances sur la surveillance et la lutte contre la haine en ligne générées dans le cadre de ce projet, ZARA a produit un manuel de formation pour permettre à d'autres acteurs de mener des formations dans ce domaine.

Perspectives futures et recommandations politiques

Au cours des deux dernières années, les partenaires de sCAN ont travaillé en étroite collaboration pour analyser et surveiller les discours de haine en ligne et pour développer des formations en ligne et hors ligne. Nous avons mis nos connaissances à la disposition du grand public et contribué à renforcer les capacités de la société civile pour combattre ensemble les discours de haine. Grâce à nos activités, nous avons recueilli de précieuses expériences et des idées d'amélioration. Tous les groupes de parties prenantes sont invités à intensifier leurs efforts pour garantir un environnement en ligne respectueux et inclusif pour tous les utilisateurs.

Le projet a fourni des recommandations politiques pour les institutions de l'Union européenne, les autorités nationales et les institutions publiques, les politiciens et les personnalités publiques, les sociétés de réseaux sociaux, les médias et les journalistes ainsi que les organisations de la société civile et les internautes individuels sur la manière de mieux combattre toutes les formes de discours haineux en ligne.

Nous recommandons à l'Union européenne d'encourager davantage de sociétés de réseaux sociaux à adhérer au code de conduite sur la lutte contre les discours de haine illégaux en ligne et de prêter attention aux petites plateformes qui peuvent être considérées comme des « refuges » pour la promotion de l'intolérance et des discours de haine en ligne. En outre, nous recommandons également de modifier la méthodologie des exercices de monitoring afin de mettre davantage l'accent sur les groupes et les comptes individuels qui diffusent constamment des discours de haine à un public important ou qui agissent comme un catalyseur de discours de haine illégaux.

Les gouvernements devraient concevoir des plans d'action nationaux pour lutter contre le discours de haine et établir ou affiner leurs systèmes nationaux de collecte de données sur ce thème, afin de garantir l'efficacité des registres des infractions pénales et des délits.

Nous invitons les responsables politiques et autres personnalités publiques à établir une responsabilité sociale politique et à s'abstenir de diffuser ou de faciliter les discours de haine en ligne. Tous les partis politiques devraient condamner les discours de haine et appeler leurs membres et leurs partisans à s'abstenir de propager la haine en ligne, y compris pendant les campagnes électorales.

Les sociétés de réseaux sociaux devraient faire plus des efforts pour appliquer efficacement leurs directives communautaires et d'encourager une communication en ligne respectueuse. En raison de leur impact important sur la société, les discours de haine diffusés par des personnalités politiques ou publiques devraient être clairement étiquetés comme tels et sanctionnés selon les normes communautaires des entreprises.

Les médias devraient veiller à fournir des reportages impartiaux sur les communautés défavorisées et renforcer leur coopération avec les OSC travaillant dans le domaine de la protection des droits de l'homme et

les représentants des communautés défavorisées afin de sensibiliser les journalistes aux stéréotypes et aux récits de discours de haine auxquels ces communautés sont couramment confrontées en ligne.

Si la plupart des OSC ont tendance à se concentrer sur des types spécifiques de discours de haine tels que le racisme, l'antisémitisme ou l'islamophobie, il est important d'inclure également d'autres types de discours de haine (par exemple, misogyne, homophobie, transphobie, interphobie, capacitisme, âgisme) dans leurs analyses et leurs campagnes d'action. En outre, les OSC devraient multiplier leurs efforts pour signaler et contrer les discours de haine intersectionnels.

Tous les internautes peuvent contribuer à freiner les discours de haine en faisant preuve de solidarité avec les personnes et les communautés visées par la haine en ligne, en contestant les expressions de haine par des contre-discours, en déconstruisant les stéréotypes haineux et en démystifiant les fausses nouvelles et les manipulations.

En ajoute aux recommandations adressées à des groupes d'acteurs spécifiques, une coopération plus étroite entre les OSC, les membres des communautés concernées, les médias, le secteur de l'internet et les pouvoirs publics est nécessaire pour freiner efficacement la diffusion des discours de haine en ligne. L'internet n'étant pas limité par les frontières nationales, une coopération transnationale accrue est fondamentale entre tous les groupes de parties prenantes pour trouver une approche commune à ce problème.

Le partenariat sCAN fera le point sur les enseignements tirés et les résultats de ses recherches des deux dernières années dans le but de planifier des initiatives de suivi visant à améliorer et à accroître sa contribution aux efforts de surveillance, d'analyse, de formation et de sensibilisation menés contre toutes les formes de discours haineux en ligne.

Ressources et lectures complémentaires

Le ressources du projet sCAN

Retrouvez tous les résultats du projet et d'autres informations sur le blog du projet:
www.scan-project.eu

sCAN Annual Report May 2018 – June 2019:
http://scan-project.eu/wp-content/uploads/sCAN_monitoring_report_year_1.pdf

sCAN Hate Ontology:
http://scan-project.eu/wp-content/uploads/scan-hate-ontology.pdf

sCAN Mapping Study "Countering online hate speech with automated monitoring tools":
http://scan-project.eu/wp-content/uploads/scan-mapping-study.pdf

User Guide on Monitoring Software:
http://scan-project.eu/wp-content/uploads/sCAN-project-Online-User-Guide.pdf

Analytical Paper "Antigypsyism on the Internet":
http://scan-project.eu/wp-content/uploads/scan-antigypsyism.pdf

Analytical Paper "Beyond the "Big Three" - Alternative platforms for online hate speech":
http://scan-project.eu/wp-content/uploads/scan-antigypsyism.pdf

Analytical Paper "Hot Spots of Hate":
http://scan-project.eu/wp-content/uploads/scan_analytical-paper-3_Hot-Spots_final.pdf

Analytical Paper "Intersectional Hate Speech Online":
http://scan-project.eu/wp-content/uploads/sCAN_intersectional_hate_final.pdf

sCAN Monitoring Report 2019:
http://scan-project.eu/wp-content/uploads/sCAN_monitoring_report_year_1.pdf

sCAN Monitoring Report 2020:
http://scan-project.eu/wp-content/uploads/sCAN_monitoring_report2_final.pdf

Policy Recommendations:
http://scan-project.eu/wp-content/uploads/sCAN_recommendations_paper_final.pdf

Online Course "Understanding and countering hate speech":
En anglais: *https://www.facingfacts.eu/courses/online-course-on-hate-speech*
En allemand: *https://www.facingfacts.eu/courses/hate-speech-was-tun*
En français: *https://www.facingfacts.eu/courses/combattre-les-discours-de-haine-sur-internet*
En italien: : *https://www.facingfacts.eu/courses/discorsi-dodio-online-riconoscerli-e-contrastarli*

Online Course "Hate Speech Moderation":
En anglais et en français:
https://www.facingfacts.eu/courses/moderating-online-hate-speech/

Advanced Monitoring Training:
https://www.youtube.com/watch?v=8t1p5fS2N8U&feature=youtu.be

Bibliographie

Fundacion Secretariado Gitano (2019). *Guide on intersectional discrimination – The case of Ro-ma women*, p. 6. Disponible sur *https://gitanos.org/upload/22/65/GUIDE_ON_INTERSECTIONAL_DISCRIMINATION_-_ROMA_WOMEN_-_FSG_33444_.pdf* (consulté le 08.04.2020).

International Network Against Cyber Hate (2018). *What is cyber hate*. Available at *http://www.inach.net/wp-content/uploads/WHAT-IS-CYBER-HATE-update.pdf* (consulté le 24.06.2019).

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

Partners

LICRA – International League against Racism and Antisemitism / France
www.licra.org

Jugendschutz.net / Germany
jugendschutz.net

CEJI-A Jewish contribution to an Inclusive Europe / Belgium
ceji.org

CESIE / Italy
cesie.org

ZARA (Zivilcourage und Anti-Rassismus-Arbeit) / Austria
zara.or.at

Human Rights House Zagreb / Croatia
humanrightshouse.org

University of Ljubljana, Faculty of Social Sciences / Slovenia
fdv.uni-lj.si

Romea / Czech Republic
www.romea.cz

Latvian Center for Human Rights / Latvia
cilvektiesibas.org.lv

RespectZone / France
www.respectzone.org

Associate partner:
International Network against Cyber Hate (INACH)
www.inach.net

Project funded by the European Union's Rights, Equality and Citizenship Programme (2014-2020)