Twitter

Réponse au questionnaire de l'Arcom sur les mesures de lutte contre la haine en ligne

31 mai 2023

Point d'étape sur les moyens mis en œuvre par les plateformes en ligne pour lutter contre la diffusion de contenus haineux en France

Merci de répondre pour chacun des services de plateforme en ligne de l'opérateur dont la fréquentation est supérieure à 10 millions de visiteurs uniques mensuels en France.

Des éléments chiffrés sont demandés pour l'année 2022. L'opérateur peut, en sus, fournir des chiffres plus détaillés. Des chiffres portant sur les premiers mois de 2023 intéresseraient également l'Arcom.

L'opérateur est également invité à fournir toute autre information qu'il estimerait utile sur les tendances observées s'agissant de ces indicateurs.

Les informations fournies par l'opérateur seront rendues publiques, sauf informations confidentielles qui doivent être clairement identifiées comme telles et dont la confidentialité doit être justifiée.

- Propos liminaires -

L'objectif de Twitter est de servir la conversation publique. La violence, le harcèlement et d'autres types de comportements similaires découragent les gens de s'exprimer et, en fin de compte, diminuent la valeur de la conversation publique mondiale. Nos règles visent à garantir que tout le monde puisse participer à la conversation publique librement et en toute sérénité.

Twitter dispose de trois catégories de règles pour la sécurité, la confidentialité et l'authenticité. Les règles et politiques de Twitter sont accessibles au public dans notre Centre d'aide, et nous veillons à ce qu'elles soient rédigées de manière facilement compréhensible. Nous veillons également à ce que notre Centre d'aide soit régulièrement mis à jour à chaque fois que nous modifions nos règles.

En outre, notre centre d'aide fournit des explications sur notre processus d'élaboration des politiques et sur notre philosophie en matière d'application des règles. La création d'une nouvelle politique ou la modification d'une politique nécessite des recherches approfondies sur les tendances en matière de comportement en ligne, l'élaboration d'un langage externe clair qui définit les attentes concernant ce qui est autorisé, et des indications sur la mise en application de nos règles à l'échelle de millions de Tweets. Nos politiques sont dynamiques et nous les révisons en permanence pour nous assurer qu'elles sont à jour, nécessaires et proportionnées.

Nous recueillons des contributions du monde entier afin de prendre en compte des perspectives diverses et variées sur la nature changeante du discours en ligne, y compris la manière dont nos règles sont appliquées et interprétées dans différents contextes culturels et sociaux. Nous testons ensuite la règle proposée avec des échantillons de Tweets potentiellement abusifs ou haineux pour mesurer l'efficacité de la politique et, une fois que nous avons déterminé qu'elle répond à nos attentes, nous élaborons et mettons en œuvre des changements de produits pour soutenir la mise à jour. Enfin, nous formons nos équipes, mettons à jour les règles de Twitter et commençons à appliquer la politique en question. L'exemple le plus récent pour illustrer ce point est notre politique de harcèlement ciblé, qui a été lancée en avril 2023 (vous pouvez également vous référer au fil de tweet de Twitter Safety).

Lorsqu'il s'agit d'appliquer nos règles, nous donnons aux gens les moyens de comprendre les différents aspects du sujet et nous encourageons les opinions et les points de vue divergents à être discutés ouvertement. Cette approche permet à de nombreuses formes de discours d'exister sur notre plateforme et, en particulier, encourage le contre-discours : c'est à dire un type de discours qui présente des faits pour corriger des inexactitudes ou des perceptions erronées, souligne les contradictions, met en garde contre les conséquences hors ligne ou en ligne, dénonce les discours haineux ou dangereux, ou contribue à faire évoluer les perceptions.

Le contexte est donc important. Pour déterminer s'il y a lieu de prendre des mesures, nous pouvons tenir compte d'un certain nombre de facteurs, y compris (mais sans s'y limiter) :

- Le comportement est dirigé contre une personne, un groupe ou une catégorie protégée de personnes
- Le rapport a été déposé par la victime ou par un témoin
- L'utilisateur a l'habitude de violer nos politiques
- La gravité de la violation
- Le contenu peut être un sujet d'intérêt public légitime.

Lorsque nous prenons des mesures à l'encontre d'un contenu, nous pouvons le faire soit sur un élément de contenu spécifique (par exemple, un Tweet ou un Message direct individuel), soit sur un compte. Il se peut que nous utilisions une combinaison de ces options. Dans certains cas, cela est dû au fait que le comportement enfreint les Règles de Twitter. Dans d'autres cas, il peut s'agir d'une réponse à une demande valide et correctement cadrée émanant d'une entité autorisée dans un pays donné.

Le 25 avril 2023, nous avons publié un billet de blog fournissant des chiffres de transparence pour le premier semestre 2022 sur les efforts de Twitter en matière de santé et de sécurité au niveau global. Au cours de la période considérée, Twitter a exigé des utilisateurs qu'ils suppriment **6 586 109 éléments de contenu** qui

violaient les Règles de Twitter, **soit une augmentation de 29 % par rapport au second semestre 2021.** Nous avons pris des mesures coercitives à l'encontre de **5 096 272 comptes** au cours de cette période (soit une augmentation de 20 % par rapport au second semestre de 2021), et **1 618 855 comptes ont été suspendus** pour avoir enfreint les Règles de Twitter **(soit une augmentation de 28 %** par rapport au second semestre de 2021).

Policy	Accounts actioned	Accounts suspended	Content removed
Abuse/Harassment	1,083,788	96,284	1,524,067
Child Sexual Exploitation	696,015	691,704	11,927
Hacked Materials	65	0	135
Hateful Conduct	1,085,651	111,056	1,527,442
Illegal or Certain Regulated Goods or Services	399,297	249,328	1,365,341
Impersonation	266,034	249,572	19,798
Misleading and Deceptive Identities	2	0	2
Non-Consensual Nudity	68,714	16,670	115,226
Perpetrators of Violent Attacks	381	0	1,578
Private Information	45,844	2,536	78,357
Promoting Suicide or Self Harm	439,555	11,776	547,377
Sensitive Media	1,315,670	150,757	1,352,155
Terrorism/Violent Extremism	30,616	30,616	0
Violence	28,753	19,838	35,240

Dans le monde entier, Twitter a reçu environ **53 000 demandes légales** de suppression de contenu de la part de gouvernements au cours de la période considérée. Twitter a reçu plus de **16 000 demandes d'informations gouvernementales pour des données d'utilisateurs provenant de plus de 85 pays** au cours de cette même période. Les taux de réponse *(disclosure rate)* varient selon le pays demandeur. Les cinq premiers pays demandeurs d'informations sur les comptes au premier semestre 2022 étaient **l'Inde, les États-Unis, la France, le Japon et l'Allemagne.**

Twitter

Réponse au questionnaire de l'Arcom sur les mesures de lutte contre la haine en ligne

31 mai 2023

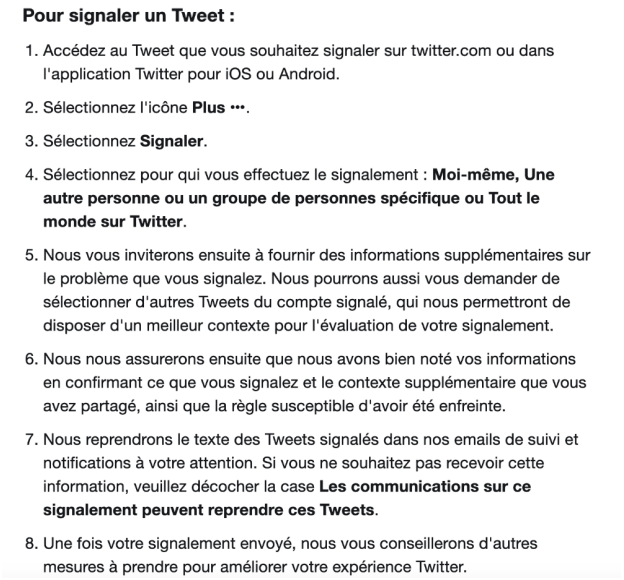
SIGNALEMENTS

1. Avez-vous rencontré des difficultés pour inclure les motifs de haine en ligne dans le dispositif de signalement mis à disposition des utilisateurs de votre service ? Merci de préciser.

Les utilisateurs de Twitter peuvent signaler des violations directement à partir d'un Tweet, d'une liste ou d'un profil individuel pour certaines violations, notamment : le spam, le contenu abusif, le harcèlement ou la haine, les publicités inappropriées, les informations privées, l'automutilation et l'usurpation d'identité. Les utilisateurs peuvent signaler un *Twitter Space* ou tout compte d'un *Twitter Space*. Ils peuvent également signaler des médias dans des Tweets *(Tweets for Media)* s'ils estiment qu'ils présentent un contenu sensible. Les utilisateurs peuvent également signaler des types de violations spécifiques, notamment les comportements abusifs et les menaces violentes, qui peuvent être signalés sur notre Centre d'Assistance par toute personne connectée ou non à Twitter.

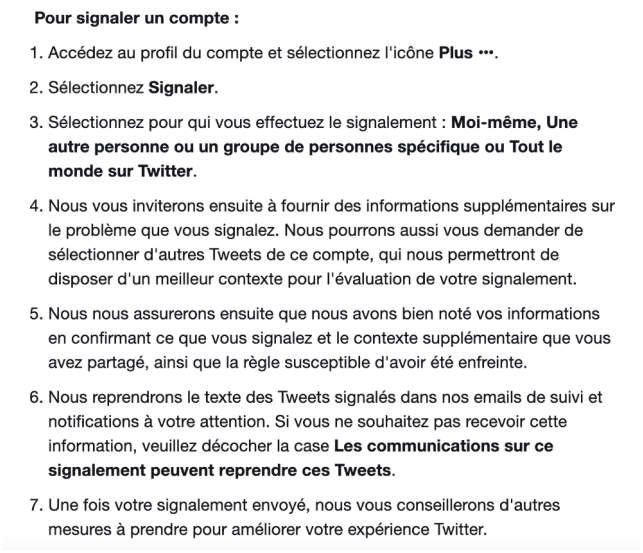
Twitter s'efforce de fournir un environnement dans lequel les gens peuvent s'exprimer librement. Si un comportement abusif se produit, nous voulons que les gens puissent facilement nous le signaler. Plusieurs Tweets peuvent être inclus dans le même rapport, ce qui nous permet d'obtenir un meilleur contexte, tout en enquêtant sur les problèmes afin de les résoudre plus rapidement. Vous trouverez dans notre centre d'aide un didacticiel vidéo montrant **les différentes étapes à suivre pour signaler un tweet susceptible d'être abusif.**

Nous avons récemment mis à jour notre expérience de signalement en adoptant une approche "symptomatique" *('symptoms first')*, en fournissant un espace pédagogique pour que les utilisateurs décrivent ce qu'il se passe avec des exemples de tweets afin de fournir autant de contexte et d'éléments que possible tout au long du processus - ce qui, nous le croyons, créera une expérience plus sûre et renforcera la confiance dans le service. Vous pouvez voir ci-dessous la procédure que les utilisateurs doivent suivre pour signaler un tweet ou un compte :



Pour signaler un Tweet :

1. Accédez au Tweet que vous souhaitez signaler sur twitter.com ou dans l'application Twitter pour iOS ou Android.
2. Sélectionnez l'icône **Plus** ···.
3. Sélectionnez **Signaler**.
4. Sélectionnez pour qui vous effectuez le signalement : **Moi-même, Une autre personne ou un groupe de personnes spécifique ou Tout le monde sur Twitter**.
5. Nous vous inviterons ensuite à fournir des informations supplémentaires sur le problème que vous signalez. Nous pourrons aussi vous demander de sélectionner d'autres Tweets du compte signalé, qui nous permettront de disposer d'un meilleur contexte pour l'évaluation de votre signalement.
6. Nous nous assurerons ensuite que nous avons bien noté vos informations en confirmant ce que vous signalez et le contexte supplémentaire que vous avez partagé, ainsi que la règle susceptible d'avoir été enfreinte.
7. Nous reprendrons le texte des Tweets signalés dans nos emails de suivi et notifications à votre attention. Si vous ne souhaitez pas recevoir cette information, veuillez décocher la case **Les communications sur ce signalement peuvent reprendre ces Tweets**.
8. Une fois votre signalement envoyé, nous vous conseillerons d'autres mesures à prendre pour améliorer votre expérience Twitter.



Pour signaler un compte :

1. Accédez au profil du compte et sélectionnez l'icône **Plus** ···.
2. Sélectionnez **Signaler**.
3. Sélectionnez pour qui vous effectuez le signalement : **Moi-même, Une autre personne ou un groupe de personnes spécifique ou Tout le monde sur Twitter**.
4. Nous vous inviterons ensuite à fournir des informations supplémentaires sur le problème que vous signalez. Nous pourrons aussi vous demander de sélectionner d'autres Tweets de ce compte, qui nous permettront de disposer d'un meilleur contexte pour l'évaluation de votre signalement.
5. Nous nous assurerons ensuite que nous avons bien noté vos informations en confirmant ce que vous signalez et le contexte supplémentaire que vous avez partagé, ainsi que la règle susceptible d'avoir été enfreinte.
6. Nous reprendrons le texte des Tweets signalés dans nos emails de suivi et notifications à votre attention. Si vous ne souhaitez pas recevoir cette information, veuillez décocher la case **Les communications sur ce signalement peuvent reprendre ces Tweets**.
7. Une fois votre signalement envoyé, nous vous conseillerons d'autres mesures à prendre pour améliorer votre expérience Twitter.

2. Parmi les signalements que vous avez reçus concernant la version française de votre service en 2022 et qui portaient sur un motif correspondant à la définition des contenus haineux au sens de l'article 6-4 la loi n° 2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique, combien d'entre eux provenaient :
    1) Des utilisateurs de la plateforme :
    2) Des signaleurs de confiance avec lesquels vous travaillez le cas échéant :
    3) Des autorités publiques (administratives ou judiciaires) :

Au cours de la période considérée, **Twitter a reçu 1 159 206 signalements d'utilisateurs et a pris des mesures sur 289 169 éléments de contenu**. Notre taux d'action basé sur la base de ces signalements est de **24,95 %.**

Twitter a reçu 242 signalements de la part de signaleurs de confiance *(Trusted Reporters)* et notre taux d'action sur ces signalements est de **66,12 %.**

Twitter peut recevoir des demandes de retrait, des demandes d'informations ou des demandes de conservation de données de la part des autorités françaises. Au cours de la période considérée, Twitter a reçu **869 demandes de suppression de contenu**, et le taux d'action de Twitter pour ces demandes était de **43,5 %.** Twitter a reçu **5032 demandes d'information**, avec un taux d'action de **42,90 %** pour ces demandes. **Nous n'avons pas reçu de demandes de conservation** au cours de la période considérée.

EXERCICE DE LA MODÉRATION

3. Combien d'actions de modération portant sur des contenus correspondant la définition des contenus haineux au sens de l'article 6-4 susmentionné avez-vous effectué en 2022, réparties par type[1] ?

Twitter a pris des **mesures sur 289 169 éléments de contenus en vertu de la LCEN** et au cours de la période de référence, ce qui correspond à un **taux d'action de 24,95 %**. Nous ne disposons pas actuellement des détails concernant les différents types d'actions de modération.

4. Combien de mesures avez-vous pris en 2022 à l'encontre d'utilisateurs abusant des outils de signalement, réparties par type ?

Dans le cadre de la mise en conformité avec le Règlement européen du DSA, Twitter se prépare aux changements de produits qui doivent être apportés à notre système de notification et d'action, notamment en cas de signalements manifestement infondés.

Dans des cas très limités, Twitter suspend des utilisateurs, par exemple en cas d'abus répété des outils de signalement. **Il n'y a pas eu de suspension d'utilisateurs dans le cadre de la LCEN pour l'année 2022.**

5. Quelle est la part des actions visées aux questions 3 d'une part, 4 d'autre part, prises après décision humaine (i.e. actions qui ne résultent pas uniquement d'un processus automatique) ?

Au cours de la période considérée, Twitter a reçu **3 405 570 signalements** (couvrant les signalements d'utilisateurs et les signaux d'apprentissage automatique *(machine learning)*). **34 %** de tous ces signalements ont fait l'objet d'une action après un examen humain.

6. À la suite des actions visées aux questions 3 d'une part, et 4 d'autre part, quel est le taux de recours internes provenant des utilisateurs accédant depuis la France à vos services, et quels ont été les résultats de ceux-ci (pourcentage de confirmation de la décision initiale, pourcentage d'infirmation de la décision initiale) ?

[1] Notamment (liste non exhaustive) : avertissement, retrait de contenu, mesures de restriction d'accès (par âge, par zone géographique), mesure de limitation de la visibilité d'un contenu ou d'un compte, mesure de démonétisation, mesures de suspension ou suppression d'un compte.

Au cours de la période considérée, Twitter a reçu **11 538 appels au titre de la LCEN**. **600 contenus** ont fait l'objet d'une action suite à ces appels, soit un taux d'action de **5,20 %.**

7. Quels sont le nombre, la localisation et la ou les langues de travail des personnes affectées au traitement des signalements et des recours provenant des utilisateurs de la version française de votre service en matière de haine en ligne (donner des informations valant au mois de décembre 2022) ?

Pour faire respecter nos règles, nous utilisons une combinaison d'apprentissage automatique et d'évaluation humaine. Nos systèmes sont capables de faire remonter les contenus vers les modérateurs humains qui utilisent le contexte pour prendre des décisions sur les violations potentielles des règles. Ce travail est mené par une équipe internationale et interfonctionnelle qui peut intervenir 24 heures sur 24 et dans plusieurs langues. Nous disposons également d'une procédure d'appel pour toute erreur potentielle. Les équipes de modération sont situées dans l'Union européenne et basées en Irlande et au Portugal, et sont composées de 149 personnes, dont des francophones.

Pour les contenus manifestement illégaux comme, par exemple, l'exploitation sexuelle des enfants et les contenus terroristes, nous nous appuyons de plus en plus sur la technologie, qui nous permet d'appliquer à grande échelle nos règles sur ce type de contenu qui n'a pas sa place sur Twitter, et ce de manière proactive.

COOPÉRATION AVEC LES AUTORITÉS PUBLIQUES

8. En 2022, combien de suspicions d'infraction ont fait l'objet d'une transmission de votre part aux autorités publiques compétentes, en particulier au ministère public, et pour quels motifs (donner, le cas échéant, le seul motif principal) :

En 2022, **Twitter n'a pas transmis de manière proactive d'infractions présumées aux autorités compétentes.** Nous avons répondu à des processus légaux valides pour des demandes d'information et de suppression.

9. Quels procédures et moyens humains et technologiques avez-vous mis en œuvre pour répondre aux autorités administratives ou judiciaire dans les meilleurs délais ?

La coopération avec les autorités chargées de l'application de la loi dans le monde entier est cruciale pour Twitter. Nous travaillons en étroite collaboration avec les forces de l'ordre du monde entier - et la France ne fait pas exception - et nous faisons de notre mieux pour les aider à identifier les utilisateurs dont le contenu est susceptible d'enfreindre les lois nationales ou locales. Toute autorité ou agence chargée de l'application de la loi peut trouver des lignes directrices sur notre Centre d'aide à destination des forces de l'ordre et peut contacter Twitter à l'aide d'un formulaire dédié.

Twitter

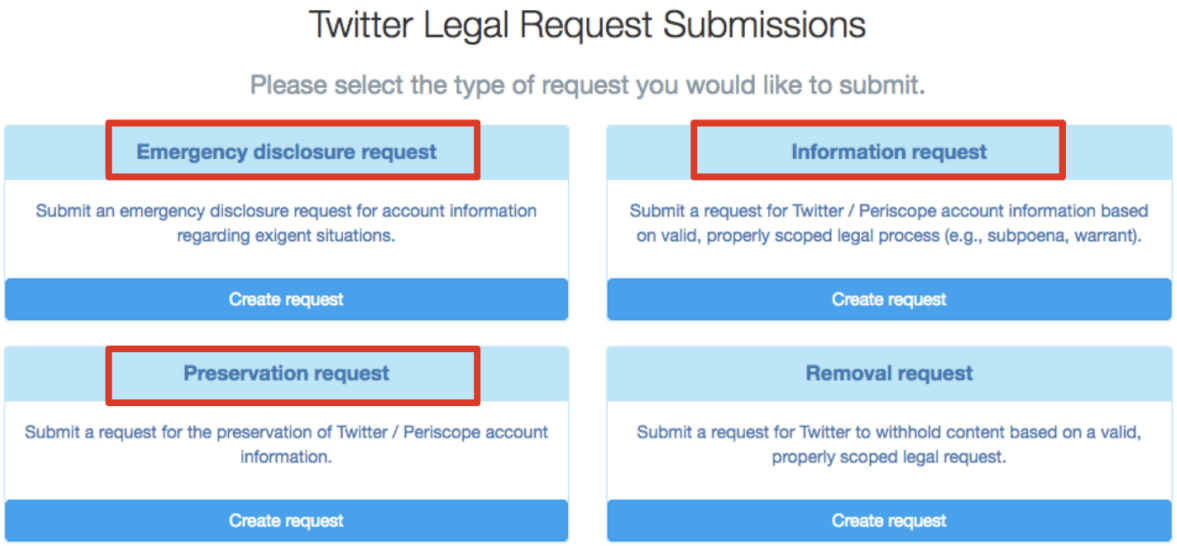
Réponse au questionnaire de l'Arcom sur les mesures de lutte contre la haine en ligne

31 mai 2023

Twitter International Unlimited Company, dont le siège se trouve à Dublin, en Irlande, examine et traite les demandes de données d'utilisateurs émanant des autorités chargées de l'application de la loi. Twitter reçoit et répond aux demandes relatives aux données des utilisateurs émanant des autorités policières et judiciaires des pays de l'Union européenne lorsqu'il existe une procédure légale valide. Nous avons mis en place un certain nombre de processus, **notamment un portail en ligne dédié à l'application de la loi**, **et des équipes d'experts couvrant l'ensemble des fuseaux horaires qui examinent les rapports et y répondent dans différentes langues.**

Les forces de l'ordre du monde entier peuvent utiliser notre portail dédié pour soumettre leurs demandes et peuvent demander les informations suivantes à Twitter :

- Demandes d'informations (RI) : demande d'informations personnelles et privées sur les utilisateurs (par exemple, les informations de base sur les abonnés).
- Demandes de retrait de contenu (RR) : demande de suppression de contenu de Twitter sur la base des conditions d'utilisation de Twitter ou des lois locales.
- Demandes de conservation : Demande de conservation de données pendant 90 jours dans le cadre d'une enquête.
- Demandes d'urgence : Processus par lequel, en cas de menace imminente pour la vie ou de préjudice corporel grave, Twitter peut divulguer des informations sur les utilisateurs sans recevoir de procédure judiciaire.



PRÉPARATION DE LA MISE EN ŒUVRE DU RSN

10. Quelles sont les actions menées par votre entreprise pour préparer le service à la mise en application du règlement européen sur les services numériques (RSN) ?

Twitter se conformera à toutes les obligations qui lui incombent en vertu du DSA et procédera aux améliorations nécessaires requises par le Règlement. Twitter communique de manière proactive et continue avec la Commission européenne en ce qui concerne ses actions préparatoires. L'entreprise est très concentrée sur son travail de mise en conformité avec le DSA et se tient prête à dialoguer avec toute autorité de régulation ayant un mandat de régulation dans le cadre du DSA en relation avec ce travail et à répondre à toute question spécifique sur ce processus.

11. Avez-vous fait face à des difficultés ou problématiques (de tous ordres : compréhension, méthode, calcul des indicateurs, modification du produit, etc.) dans la mise en œuvre de l'article 6-4 susmentionné qui seraient susceptibles d'être de nouveau rencontrées dans le cadre de celle du RSN, ou qui vous auraient permis d'anticiper ce dernier ? Lesquelles ?

Twitter n'a pas de commentaire spécifique à faire sur les difficultés ou les problèmes rencontrés dans la mise en œuvre de l'article 6-4 susmentionné.

12. Travaillez-vous avec des tiers établis en France que vous reconnaissez comme signaleurs de confiance en matière de haine en ligne ? Si oui, lesquels ? Le cas échéant, pour quelles raisons avez-vous choisi de collaborer avec eux ?

Twitter a travaillé avec les partenaires de confiance suivants en France : 3018/e-enfance, Point de Contact, la Licra, SOS Homophobie et Le CRIF. Twitter définit les Signaleurs de Confiance *(Trusted Reporters)* comme des organisations dont la mission consiste à signaler à des fournisseurs de services tels que Twitter des contenus susceptibles d'être considérés comme des comportements haineux au regard de la législation européenne, tels que définis dans le Code de conduite de l'Union européenne sur les discours de haine illégaux.

Nous disposons d'un canal de traitement prioritaire pour les signalements émanant de nos partenaires. Il s'agit d'un portail avec un accès dédié qui vous permet de signaler un ou plusieurs Tweets et/ou comptes. Les signalements reçus sont traités en priorité par nos équipes en charge de la modération.