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.

Introduction

LinkedIn est un réseau social en ligne à l'identité réelle permettant aux professionnels de se connecter et d'interagir avec d'autres professionnels, de développer leur réseau et leur image, et de rechercher des opportunités de développement de carrière.

LinkedIn fait partie de l'identité professionnelle de ses membres et a un objectif spécifique. L'activité sur la plateforme et le contenu partagé par les membres peuvent être vus par les employeurs actuels et futurs, les collègues, les partenaires commerciaux potentiels et les cabinets de recrutement, entre autres. Compte tenu de cette audience, les membres limitent largement leur activité à des domaines d'intérêt professionnel et s'attendent à ce que le contenu qu'ils voient soit de nature professionnelle.

LinkedIn s'engage à assurer la sécurité, la confiance et le professionnalisme de sa plateforme et respecte les lois qui s'appliquent à ses services. En rejoignant LinkedIn, les membres acceptent de se conformer aux Conditions générales d'utilisation de LinkedIn et à ses Politiques de la communauté professionnelle, qui interdisent expressément aux membres de publier des propos haineux. Plus précisément, les règles de la communauté professionnelle prévoient ce qui suit:

> **Ne tenez pas de propos haineux:** nous interdisons tout contenu qui attaque, dénigre, intimide, déshumanise, menace de ou incite à des actes de haine, de violence, discriminatoires ou attentatoires contre des individus ou des groupes en raison de leur appartenance ethnique réelle ou supposée, leur couleur de peau, leur origine nationale, leur sexe, leur identité de genre, leur orientation sexuelle, leur affiliation religieuse, leur âge ou leur statut de handicap. Les groupes haineux ne sont pas

autorisés sur LinkedIn. N'utilisez sous aucun prétexte les insultes racistes, religieuses ou autres qui incitent à ou promeuvent la haine, ni tout autre contenu destiné à créer la division. Ne publiez pas et ne partagez pas de contenu niant un événement historique documenté, comme l'Holocauste ou l'esclavage aux États-Unis.

En savoir plus sur nos politiques en matière de contenu haineux et dénigrant.

Le lien ci-dessus pour « En savoir plus » fournit plus de détails sur nos politiques ainsi que des exemples de types de contenus haineux et désobligeants.

Lorsque LinkedIn décèle un contenu ou un comportement qui enfreint ses Politiques de la communauté professionnelle, il prend des mesures, notamment la suppression du contenu ou la restriction d'un compte en cas de comportement abusif répété.

LinkedIn utilise une approche multidimensionnelle à trois niveaux pour modérer le contenu au sein de son écosystème de confiance.

Première couche de protection - Prévention automatique et proactive
Lorsqu'un membre tente de créer un contenu sur LinkedIn, divers appels (ou signaux) sont envoyés aux services d'apprentissage automatique de LinkedIn. Ces services visent à filtrer automatiquement certains contenus violant les politiques dans les 300 millisecondes suivant leur création, ce qui signifie que le contenu n'est visible que par l'auteur et n'est montré à personne d'autre sur la plateforme. Dans le cadre de ce processus, l'intelligence artificielle (IA) joue un rôle clé en aidant LinkedIn à filtrer de manière proactive les contenus potentiellement nuisibles. LinkedIn utilise du contenu (comme certains mots clés ou images) dont il a déjà été établi qu'il viole ses Politiques de la communauté professionnelle pour aider à informer les modèles d'IA et mieux identifier et restreindre la publication d'un contenu similaire à l'avenir.

LinkedIn évalue régulièrement ses services de défense préventive afin d'améliorer la précision du processus de filtrage. Pour ce faire, il envoie des échantillons positifs à des fins d'examen humain afin de mesurer la précision du système de défense automatisé de LinkedIn. Cela réduit la probabilité que le processus de filtrage automatique de LinkedIn supprime des contenus conformes aux politiques de LinkedIn.

Deuxième niveau de protection - Combinaison de détection automatique et humaine
Le deuxième niveau de modération de LinkedIn détecte les contenus susceptibles d'être en infraction, mais pour lesquels l'algorithme n'est pas suffisamment sûr pour justifier une suppression automatique. Ce contenu est signalé par nos systèmes d'IA pour un examen humain plus approfondi. Si l'équipe de révision humaine détermine que le contenu viole les politiques de LinkedIn, il est retiré de la plateforme. L'équipe d'évaluation humaine de LinkedIn joue un rôle essentiel dans ce processus et aide à former les modèles de la plateforme.

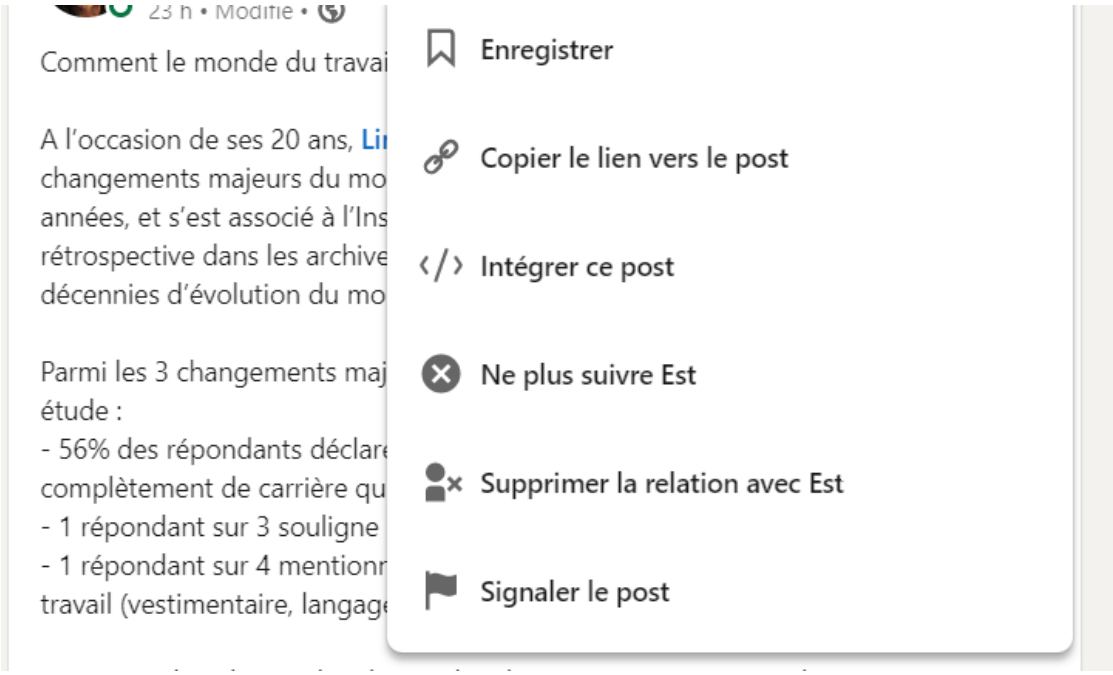
Troisième niveau de protection – Détection humaine
Si les membres trouvent un contenu qui, selon eux, enfreint les Politiques de la communauté professionnelle, nous les encourageons à le signaler en utilisant le mécanisme de signalement intégré, représenté par les trois points dans le coin supérieur droit du contenu lui-même sur LinkedIn. Le contenu signalé est ensuite envoyé à l'équipe de réviseurs de LinkedIn pour une évaluation plus approfondie et est supprimé s'il est jugé en violation des politiques de LinkedIn.

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.

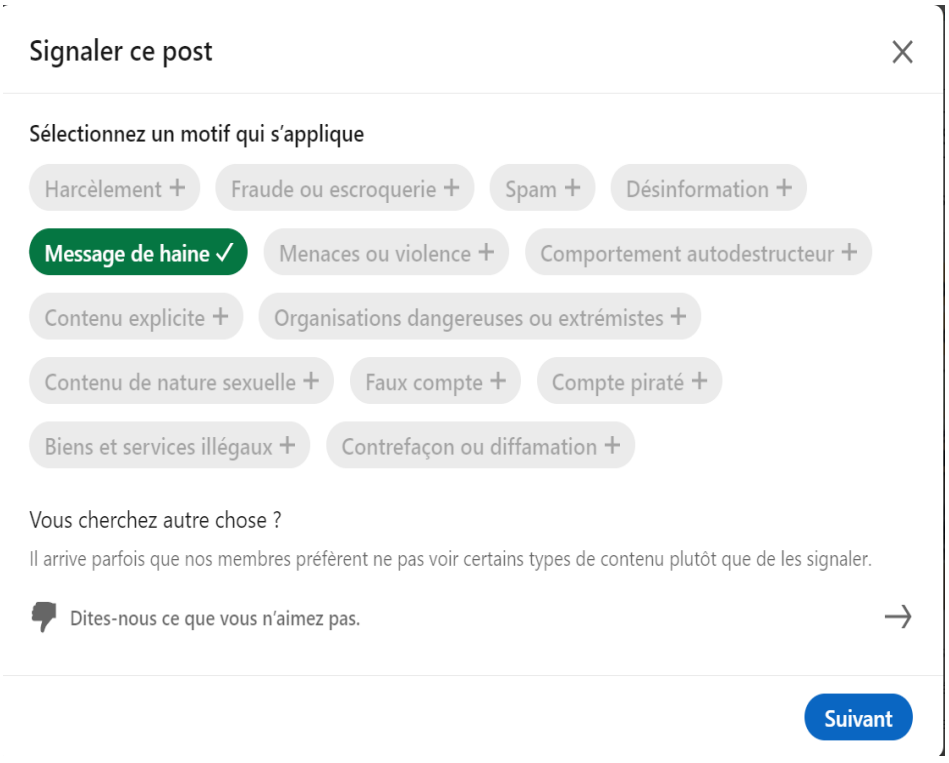
   Non, si les membres de LinkedIn trouvent du contenu qui, selon eux, enfreint les Politiques de la communauté professionnelle, LinkedIn les encourage à le signaler en utilisant le mécanisme de signalement intégré, représenté par les trois points dans le coin supérieur droit du contenu lui-même sur LinkedIn.

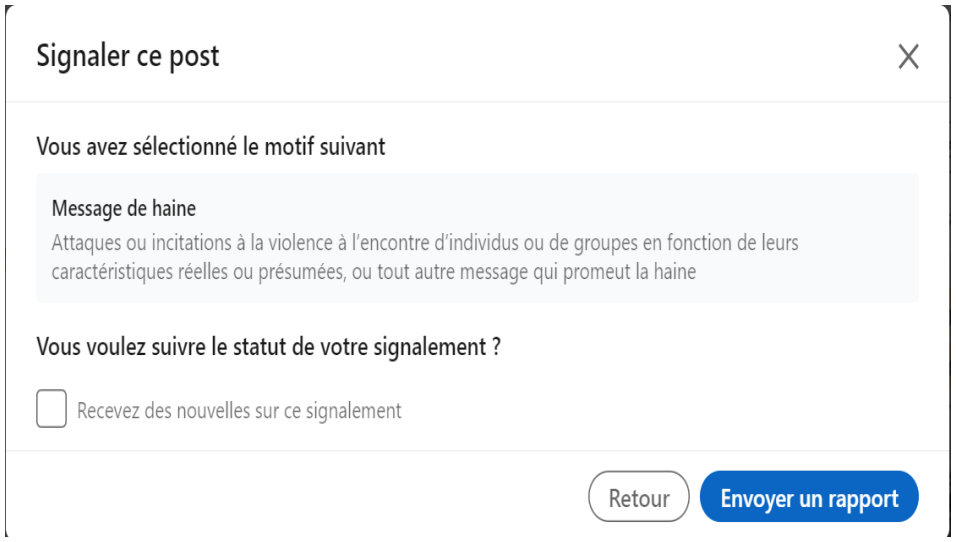
   Les discours haineux sont expressément mentionnés comme l'une des options de signalement.[1]



1 LinkedIn est en train de déployer une version actualisée de son flux de signalement. Les captures d'écran ci-dessous reflètent ce nouveau flux, qui est actuellement disponible pour les membres situés en France pour certains types de contenu. Toutefois, les discours haineux étaient également mentionnés comme motif de signalement potentiel dans la version précédente du flux, qui était en place en 2022.


Signaler ce post
Sélectionnez un motif qui s'applique
Harcèlement +
Fraude ou escroquerie +
Spam +
Désinformation +
Message de haine +
Menaces ou violence +
Comportement autodestructeur +
Contenu explicite +
Organisations dangereuses ou extrémistes +
Contenu de nature sexuelle +
Faux compte +
Compte piraté +
Biens et services illégaux +
Contrefaçon ou diffamation +
Vous cherchez autre chose ?
Il arrive parfois que nos membres préfèrent ne pas voir certains types de contenu plutôt que de les signaler.
Dites-nous ce que vous n'aimez pas.
Suivant


Signaler ce post
Sélectionnez un motif qui s'applique
Harcèlement +
Fraude ou escroquerie +
Spam +
Désinformation +
Message de haine ✓
Menaces ou violence +
Comportement autodestructeur +
Contenu explicite +
Organisations dangereuses ou extrémistes +
Contenu de nature sexuelle +
Faux compte +
Compte piraté +
Biens et services illégaux +
Contrefaçon ou diffamation +
Vous cherchez autre chose ?
Il arrive parfois que nos membres préfèrent ne pas voir certains types de contenu plutôt que de les signaler.
Dites-nous ce que vous n'aimez pas.
Suivant



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 allant du 1er janvier 2022 au 31 décembre 2022-

    - [✂] contenus ont été signalés comme discours haineux par des membres français (sur la base de l'emplacement du profil du membre).
    - LinkedIn n'a reçu aucune injonction de la part des autorités publiques françaises.

    En ce qui concerne le point 2), voir la réponse de LinkedIn à la question 12 ci-dessous.

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[2]?

    Lorsqu'un contenu est identifié comme violant nos politiques (que ce soit à la suite d'un signalement ou détecté de manière proactive), nous ne l'étiquetons pas, mais le retirons plutôt de LinkedIn. Au cours de la période du 1er janvier 2022 au 31

2 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.

décembre 2022, 3 064 contenus publiés ou transmis par des membres français (sur la base de l'emplacement du profil du membre) ont été supprimés au motif que ces contenus violaient les politiques en matière de contenu haineux et dénigrant de LinkedIn.

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

   Ce n'est pas quelque chose qui a fait l'objet d'un suivi historique et, par conséquent, LinkedIn n'a pas de mesures à rapporter pour 2022. Cependant, la loi sur les services numériques comprend des obligations visant à traiter les abus des outils de signalement [par exemple, l'article 23, paragraphe 2, exige que LinkedIn prenne certaines mesures en réponse à la soumission fréquente par un individu ou une entité d'avis et/ou de plaintes manifestement non fondés. L'article 24, paragraphe 1, point b), prévoit des obligations de signalement connexes]. Voir la réponse à la question 10 pour plus d'informations sur le travail de LinkedIn visant à mettre en œuvre la loi sur les services numériques.

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)?

   Sur les 3 064 actions mentionnées à la question 3 ci-dessus, LinkedIn a supprimé [✂] contenus à la suite d'un examen humain et pas uniquement par des moyens automatisés.

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)?

   Sur les 3 064 actions mentionnées à la question 3 ci-dessus–
   - 5 ont fait l'objet d'un recours; et
   - sur ces recours;
     - dans les 5 cas, LinkedIn a confirmé sa décision initiale; et
     - dans aucun cas LinkedIn n'est revenu sur sa décision initiale.

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)?

   LinkedIn dispose d'une équipe interne composée de centaines de réviseurs de contenu situés dans le monde entier (pour une couverture 24/7) avec une expertise linguistique locale (dont environ 29 spécialistes de la langue française). Par exemple, fin 2022, l'équipe de révision comprenait un total combiné d'environ 278 employés à

temps plein basés dans la région EMEA et des contractants qui se consacraient à la modération du contenu.

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) :

   LinkedIn n'a reçu aucune saisine de ce type de la part des autorités publiques compétentes françaises au cours de la période allant du 1er janvier 2022 au 31 décembre 2022.

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

   LinkedIn a créé des canaux spécialisés pour que les autorités puissent nous contacter dans de telles situations. Le personnel de LinkedIn surveille ces canaux en continu afin de les évaluer et d'y répondre rapidement et en conséquence.

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) ?

    Le 25 avril 2023, LinkedIn Ireland Unlimited Company (le prestataire des services de LinkedIn dans l'UE, basé à Dublin) a été désigné comme très grande plateforme en ligne conformément à l'article 22, paragraphe 4, de la loi sur les services numériques. Pour respecter son délai de conformité du 25 août 2023, LinkedIn a consacré et continue de consacrer des ressources importantes à la mise en œuvre de ses obligations en vertu de la loi sur les services numériques. Dans le cadre de ces efforts, LinkedIn s'est également engagé régulièrement auprès de la Commission européenne et de la Coimisiún na Meán, le régulateur qui devrait être désigné comme coordinateur des services numériques pour l'Irlande (l'État membre dans lequel LinkedIn Ireland Unlimited Company a son principal établissement) aux fins de la loi sur les services numériques.

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?

    À ce jour, pas à notre connaissance.

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?

    LinkedIn n'a pas actuellement de programme formel pour les signaleurs de confiance. Cependant, LinkedIn s'engage auprès des signaleurs de confiance sur une base ad hoc et encourage les membres à signaler les contenus abusifs, les messages ou les problèmes de sécurité, que ce soit dans les profils, les communications, les messages, les commentaires ou n'importe où ailleurs. LinkedIn est également signataire du code de conduite de l'UE sur la lutte contre les discours haineux illégaux en ligne depuis 2021. Dans le cadre de la mise en œuvre par LinkedIn de la loi sur les services numériques, LinkedIn s'engagera auprès des entités ayant obtenu le statut de signaleur de confiance conformément à l'article 22, paragraphe 2.