

CAMERA DEI DEPUTATI

Doc. **XII-bis**
n. **189**

ASSEMBLEA PARLAMENTARE DEL CONSIGLIO D'EUROPA

Risoluzione n. 2590

Regolamentare la moderazione dei contenuti sui *social media*
per salvaguardare la libertà di espressione

Trasmessa il 4 febbraio 2025

PARLIAMENTARY ASSEMBLY OF THE COUNCIL OF EUROPE

RESOLUTION 2590 (2025) ⁽¹⁾

Provisional version

Regulating content moderation on social media to safeguard freedom of expression

PARLIAMENTARY ASSEMBLY

1. Social media have become an online agora where users come to exercise their right to freedom of expression and information in many ways. These include posting their own content and enjoying the content posted by others, getting informed and informing others, and communicating with other users.

2. The right to freedom of expression is not an absolute right; social media are legally obliged to remove any illegal content when they become or are made aware of its existence on their services. Moreover, it is incumbent upon social media to combat the dissemination of harmful content.

3. Social media companies are also bearers of fundamental rights, such as the right to property and freedom of enterprise, and therefore they have a say in how users can use their services and what content they can post. The content moderation rules included in their terms and conditions allow for social media companies to demote, demonetise, restrict access to, or remove a concrete content item because of

its incompatibility with their terms and conditions. In extreme cases, social media companies can suspend or even terminate a user's account. Their terms and conditions have a contractual character, and users are bound by them on a take-it-or-leave-it basis.

4. The major social media companies, mainly US owned, have a global reach; their content moderation policies and their commercial or ideological decisions about content to promote or demote may have an immense influence on public opinion and on choices of billions of people. It is, nevertheless, incumbent upon them to respect the laws of the country in which they provide their services.

5. Given the potential impact on societal behaviours and on the proper functioning of democratic processes that the information and communication flow on social media *de facto* has, it is incumbent upon the State to establish the fundamental principles and institutional framework that may correct the power imbalance resulting from the unequal contractual relationship and ensure the effective protection of the right to freedom of expression.

6. It is imperative, however, that public regulation of content moderation does not have a chilling effect on free speech and is not intended to impose the views of the

(1) *Assembly debate* on 30 January 2025 (8th sitting) (see Doc. 16089, report of the Committee on Culture, Science, Education and Media, rapporteur: Ms Valentina Grippo). *Text adopted by the Assembly* on 30 January 2025 (8th sitting).

political power in place and censorship on opinions or ideas which may conflict with the ruling majority's vested interests. Moreover, national regulations should not place undue burdens on social media, which could result in an overzealous approach to content removal. These regulations and their implementation must uphold freedom of expression and carefully assess the necessity of any restrictions.

7. The risk of restrictive content moderation policies is increased by the lack of transparency in their implementation. Social media have been accused of a practice called « shadow banning » whereby they delist or demote content dealing with controversial issues without notifying the user in question, making that content invisible to other users. This devious, hidden practice should be forbidden: it deprives users of the possibility to defend effectively their right to freedom of expression.

8. The press and the media in general use social media as a platform for disseminating information to the public. It is therefore essential that content moderation practices do not unduly impact media and journalistic content that respects professional standards and the national regulatory framework.

9. Content moderation is increasingly carried out by automated means. Artificial intelligence tools are much more efficient than human moderators in processing at a high speed the colossal amount of content circulating on the web, to identify prohibited content. They lack to date, however, the capacity to fully understand the subtleties of human interaction (humour, parody, satire, etc.) and to assess the content in its context.

10. For this reason, human moderators must remain the cornerstone of any content moderation system and be responsible for making decisions in cases where automated systems are not up to the task. However, human moderation can be biased and lead to inconsistencies among countries due to cultural differences; it is therefore imperative to establish clear and comprehensive standards and to guarantee appropriate training, to ensure that all modera-

tors have the requisite knowledge of both the applicable legislation and the company's internal guidelines, as well as of the language and the context of the country from which the content originates. However, in the event of a military conflict between two countries, moderators from one country party to the conflict should not moderate content originating from the other.

11. Regrettably, despite their fundamental role, human moderators' working conditions are inadequate, they are overexposed to disturbing content that can cause them serious mental health problems and they suffer from restrictions on their freedom to speak out about the problems they encounter at work.

12. Generative artificial intelligence tools allow to produce synthetic content that is virtually indistinguishable from human generated content. Such content can be highly misleading, be a tool of disinformation and manipulation, and instigate hatred and discrimination, among other dangers. It is essential that users are made aware of content that appears to be genuine, but which is in fact not. In this regard, watermarking techniques are particularly beneficial but have several drawbacks, including their lack of interoperability among social media services.

13. Independent assessment of terms and conditions and content moderation policies and their enforcement, also with a view to identifying and promoting best practices, could help to ensure their consistency with principles which uphold a human rights-based approach to content moderation.

14. The establishment of clear and transparent rules for conflict resolution is essential to ensure the protection of users and to minimise the risk of being subjected to a potentially biased decision by the social media company, or of being forced to pursue costly legal action against a multinational corporation with enormous financial resources at its disposal.

15. The establishment of independent out-of-court dispute settlement bodies to assess content moderation decisions may prove beneficial in enhancing compliance

with fundamental rights. Collaboration between social media companies in establishing such bodies could also hopefully facilitate dispute resolution.

16. As stated by the Assembly in its Resolution 2281 (2019) « Social media: social threads or threats to human rights ? », social media companies should employ algorithms that promote the diversity of sources, topics and views, guarantee the quality of information available, and thereby reduce the risk of « filter bubbles » and « echo chambers ».

17. In light of these considerations, the Assembly calls on the Council of Europe member States to review their legislation to better safeguard the right to freedom of expression on social media. In this respect, they should in particular:

17.1. require that social media uphold users' fundamental rights, including freedom of expression, in their content moderation policy and implementation practices;

17.2. require that social media platforms provide justification for any measure taken to moderate content provided by the press or media service providers prior to its implementation and allow them an opportunity to reply within an appropriate time-frame;

17.3. in co-operation with the press or media organisations, implement a system of verification of media and journalist accounts, together with robust mechanisms to protect them from online harassment, hacks and fraud, and develop social media guidelines for press or media organisations on the publication of information on sensitive issues, with a view to avoiding unnecessary moderation restrictions on this type of content;

17.4. provide for minimum standards for the working conditions of human moderators, including a requirement of adequate training to carry out their often stressful tasks and of access to proper psychological support and mental healthcare when needed;

17.5. sign and ratify the Council of Europe Framework Convention on Artifi-

cial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225, « Vilnius Convention ») and adopt or maintain measures to ensure that adequate transparency and oversight requirements tailored to the specific contexts and risks are in place to meet the challenges of the identification of content generated by artificial intelligence systems;

17.6. require that content generated by artificial intelligence is disclosed as such by those initially posting it and that social media implement technical solutions allowing for such content to be easily identified by users, and encourage collaboration between social media companies to ensure the interoperability of watermarking techniques for content generated by artificial intelligence;

17.7. require that out-of-court dispute settlement bodies, when established, are independent and impartial, have the necessary expertise, are easily accessible, and operate according to clear and fair rules, with certification of these requirements by the competent national regulatory authority;

17.8. promote, within the Internet Governance Forum and the European Dialogue on Internet Governance, reflection on the possibility for the internet community to develop, through a collaborative and, where appropriate, multi-stakeholder process, an external evaluation and auditing system aimed at determining whether algorithms are unbiased and respect the right to freedom of expression, and a « seal of good practices » which could be awarded to social media whose algorithms are designed to reduce the risk of « filter bubbles » and « echo chambers » and to foster an ideologically cross-cutting, yet safe, user experience.

18. The Assembly calls on social media companies to avoid measures that unnecessarily restrict the freedom of expression of users. They should, in particular:

18.1. directly incorporate principles of fundamental rights law, and in particular

freedom of expression, into their terms and conditions;

18.2. use caution when moderating content that is not obviously illegal;

18.3. provide users with terms and conditions that are readily accessible, clear and informative on the types of content that are permissible on their services and the consequences for non-compliance, and which are understandable to the wide span of users notwithstanding differing levels of digital literacy and reading proficiency;

18.4. notify users without undue delay of any moderation action taken on their content, providing a comprehensive account of the rationale behind the decision, accompanied by a reference to the internal rules which have been applied;

18.5. refrain from shadow banning users' content and notify users of every instance of demotion or delisting;

18.6. ensure that automated content moderation processes are subject to human oversight and to rigorous and continuous evaluation to assess their performance;

18.7. make available a system for handling complaints that is easily accessible, user-friendly, and allows users to make a precise complaint;

18.8. give human moderators appropriate training and working conditions which pay attention to the heavy psychological stress they are submitted to, and ensure adequate protection of their health;

18.9. refrain from permanent deletion of content (including its metadata) that has been removed in accordance with legal

obligations or with terms and conditions, in particular when the content in question may serve as evidence of war or other crimes;

18.10. ensure that the artificial intelligence systems they develop or use uphold Council of Europe standards, including the new Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law; algorithms should be designed to respect the right to freedom of expression, and to encourage plurality and diversity of views and opinions while ensuring a safe user experience, their operation modalities should be disclosed and users duly informed on how these algorithms filter and promote content;

18.11. collaborate with other online services with the aim of ensuring the interoperability of watermarking techniques for content generated by artificial intelligence;

18.12. collaborate with journalists and fact-checking organisations to effectively combat disinformation with information that adheres to the ethical and professional standards of journalism;

18.13. promote and support the creation of independent out-of-court dispute settlement bodies, and abide by their decisions and recommendations;

18.14. support the work of independent third-party oversight bodies and abide by their decisions and recommendations;

18.15. ensure that decisions related to content moderation are duly motivated and that researchers have access to full information on the legal base and reasoning behind each decision.

ASSEMBLÉE PARLEMENTAIRE DU CONSEIL DE L'EUROPE

RÉSOLUTION 2590 (2025) ⁽¹⁾

Version provisoire

Réglementer la modération de contenu sur les réseaux sociaux pour sauvegarder la liberté d'expression

ASSEMBLÉE PARLEMENTAIRE

1. Les réseaux sociaux sont devenus un espace en ligne où les utilisateurs et utilisatrices viennent exercer leur droit à la liberté d'expression et d'information de nombreuses manières. Ces personnes peuvent notamment publier leurs propres contenus, apprécier ceux qui sont publiés par d'autres, s'informer et informer autrui, et communiquer avec d'autres utilisateurs et utilisatrices.

2. Le droit à la liberté d'expression n'étant pas un droit absolu, les réseaux sociaux sont légalement tenus de supprimer tout contenu illégal lorsqu'ils apprennent son existence sur leurs services ou en prennent connaissance. Il leur incombe également de lutter contre la diffusion de contenus préjudiciables.

3. Les entreprises de réseaux sociaux sont également porteuses de droits fondamentaux, notamment le droit de propriété et la liberté d'entreprise; elles ont donc leur mot à dire sur la manière dont les utilisateurs et utilisatrices peuvent utiliser leurs services et sur les contenus qu'ils et elles peuvent publier. Les règles de modération

de contenu qui sont incluses dans leurs conditions générales permettent aux entreprises de réseaux sociaux de déclasser ou de démonétiser un contenu précis, d'en restreindre l'accès ou de le supprimer en raison de son incompatibilité avec ces conditions. Dans des cas extrêmes, les entreprises de réseaux sociaux peuvent suspendre, voire supprimer, le compte d'un utilisateur. Leurs conditions générales ont un caractère contractuel et sont « à prendre ou à laisser » pour l'utilisateur ou l'utilisatrice.

4. Les grandes entreprises de réseaux sociaux, qui sont principalement américaines, sont d'envergure mondiale; leurs politiques de modération de contenu et leurs décisions commerciales ou idéologiques concernant les contenus à promouvoir ou à déclasser peuvent exercer une influence considérable sur l'opinion publique et les choix de milliards de personnes. Il leur incombe néanmoins de respecter les lois du pays dans lequel elles fournissent leurs services.

5. Compte tenu de l'impact que les flux d'information et de communication qui circulent sur les réseaux sociaux peuvent avoir de fait sur les comportements sociétaux et le bon fonctionnement des processus démocratiques, il incombe à l'État de mettre en place les principes fondamentaux et le cadre institutionnel susceptibles de corriger

(1) *Discussion par l'Assemblée le 30 janvier 2025 (8^e séance) (voir Doc. 16089, rapport de la commission de la culture, de la science, de l'éducation et des médias, rapporteure: M^{me} Valentina Grippo). Texte adopté par l'Assemblée le 30 janvier 2025 (8^e séance).*

le déséquilibre des pouvoirs qui découle de cette relation contractuelle inégale et d'assurer la protection effective du droit à la liberté d'expression.

6. Il est cependant indispensable que la réglementation publique de la modération de contenu n'ait pas d'effet dissuasif sur la liberté d'expression et n'ait pas pour objet d'imposer les vues du pouvoir politique en place et de censurer les opinions ou les idées susceptibles d'entrer en conflit avec les intérêts acquis de la majorité au pouvoir. En outre, les réglementations nationales ne doivent pas imposer des contraintes excessives aux réseaux sociaux qui pourraient se traduire par un excès de zèle dans la suppression de contenus. Ces réglementations et leur mise en œuvre doivent respecter la liberté d'expression et évaluer soigneusement la nécessité de toute restriction.

7. Le risque que représentent les politiques de modération de contenu restrictives est accru par le manque de transparence de leur mise en œuvre. Les réseaux sociaux ont été accusés d'utiliser le « bannissement furtif » (en anglais *shadow banning*), une pratique qui consiste à déréférencer ou déclasser des contenus qui traitent de questions controversées sans en informer l'utilisateur ou l'utilisatrice concerné-e, ce qui les rend invisibles pour les autres utilisateurs et utilisatrices. Cette pratique sournoise et cachée doit être interdite: elle prive les utilisateurs et utilisatrices de la possibilité de défendre efficacement leur droit à la liberté d'expression.

8. La presse et les médias en général utilisent les médias sociaux comme plateforme de diffusion d'informations au public. Il est donc essentiel que les pratiques de modération des contenus n'affectent pas indûment les contenus médiatiques et journalistiques qui respectent les règles professionnelles et le cadre réglementaire national.

9. La modération de contenu est de plus en plus souvent effectuée par des moyens automatisés. Les programmes d'intelligence artificielle sont beaucoup plus efficaces que les modérateurs et modératrices humains pour traiter à grande vitesse la quantité

colossale de contenus circulant sur internet et identifier les contenus interdits. Il leur manque cependant, pour l'instant, la capacité de comprendre pleinement les subtilités des échanges humains (humour, parodie, satire, etc.) et d'évaluer le contenu dans son contexte.

10. C'est pourquoi les modérateurs et modératrices humains doivent rester la pierre angulaire de tout système de modération de contenu et être chargés de prendre des décisions dans les cas où les systèmes automatisés ne sont pas à la hauteur de la tâche. Cependant, la modération humaine peut être biaisée et entraîner des incohérences entre les pays en raison des différences culturelles. Il est donc impératif d'établir des normes claires et exhaustives et de garantir une formation appropriée, afin de s'assurer que tous les modérateurs et modératrices ont les connaissances requises à la fois de la législation applicable et des lignes directrices internes de l'entreprise, ainsi que de la langue et du contexte du pays d'où provient le contenu. Toutefois, en cas de conflit militaire entre deux pays, les modérateurs et modératrices d'un pays partie au conflit ne devraient pas modérer les contenus provenant de l'autre pays.

11. Malheureusement, les conditions de travail des modérateurs et modératrices humains sont souvent inadaptées malgré le rôle essentiel. Ces personnes sont surexposées à des contenus dérangeants qui peuvent leur causer de graves problèmes de santé mentale, et elles subissent des restrictions à leur liberté de s'exprimer sur les problèmes qu'elles rencontrent au travail.

12. Les outils d'intelligence artificielle générative permettent de produire des contenus synthétiques qui sont pratiquement impossibles à distinguer des contenus générés par un être humain. Ces contenus peuvent être très trompeurs, servir de moyen de désinformation et de manipulation, et inciter à la haine et à la discrimination, entre autres dangers. Il est essentiel que les utilisateurs et utilisatrices sachent que des contenus qui semblent authentiques ne le sont pas. Les techniques de tatouage numérique sont particulièrement bénéfiques à cet égard, même si elles présentent des in-

convénients, notamment leur manque d'interopérabilité entre les services de réseaux sociaux.

13. Une évaluation indépendante des conditions générales, des politiques de modération de contenu et de leur mise en œuvre réalisée dans le but de définir et de promouvoir de bonnes pratiques pourrait contribuer à garantir leur conformité avec les principes qui défendent une approche de la modération de contenu fondée sur les droits humains.

14. Il est essentiel d'établir des règles claires et transparentes de règlement des conflits afin d'assurer la protection des utilisateurs et utilisatrices et d'atténuer le risque d'être soumis à une décision de l'entreprise de réseaux sociaux qui pourrait être tendancieuse, ou d'être contraint d'initier une action en justice coûteuse contre une multinationale disposant d'énormes ressources financières.

15. La mise en place d'organes indépendants de règlement extrajudiciaire des litiges chargés d'évaluer les décisions de modération de contenu peut s'avérer judicieuse pour renforcer le respect des droits fondamentaux. La collaboration entre les entreprises de réseaux sociaux pour la mise en place de tels organes pourrait également faciliter le règlement des litiges.

16. Comme l'a indiqué l'Assemblée parlementaire dans sa Résolution 2281 (2019) « Médias sociaux: créateurs de liens sociaux ou menaces pour les droits humains ? », les entreprises de médias sociaux devraient utiliser des algorithmes qui favorisent la diversité des sources, des sujets et des points de vue, garantir la qualité des informations disponibles et réduire ainsi les risques de « bulles de filtrage » et de « chambres d'écho ».

17. Au vu de ces éléments, l'Assemblée appelle les États membres du Conseil de l'Europe à réviser leur législation afin de mieux protéger le droit à la liberté d'expression sur les réseaux sociaux. À cet égard, ils devraient en particulier:

17.1. exiger que les réseaux sociaux respectent les droits fondamentaux des utilisateurs et utilisatrices, notamment la liberté d'expression, dans leur politique de

modération de contenu et leurs pratiques de mise en œuvre;

17.2. exiger que les plateformes de médias sociaux justifient toute mesure prise pour modérer les contenus fournis par la presse ou les fournisseurs de services de médias avant sa mise en œuvre, et leur donner la possibilité de répondre dans un délai approprié;

17.3. en coopération avec les organes de presse ou organisations de médias, mettre en œuvre un système de vérification des comptes des médias et des journalistes, ainsi que des mécanismes solides pour les protéger contre le harcèlement en ligne, les piratages et la fraude, et élaborer des lignes directrices en matière de médias sociaux à l'attention des organes de presse ou organisations de médias sur la publication d'informations concernant des questions sensibles, en vue d'éviter d'inutiles restrictions par modération pour ce type de contenu;

17.4. prévoir des normes minimales encadrant les conditions de travail des modérateurs et modératrices humains, y compris l'exigence d'une formation adéquate pour mener à bien leurs tâches souvent stressantes et l'accès à un soutien psychologique approprié et à des soins de santé mentale en cas de besoin;

17.5. signer et ratifier la Convention-cadre du Conseil de l'Europe sur l'intelligence artificielle et les droits de l'homme, la démocratie et l'État de droit (STCE n° 225, « Convention de Vilnius ») et adopter ou maintenir des mesures pour veiller à la mise en place d'obligations de transparence et de contrôle adaptées aux contextes et aux risques spécifiques pour relever les défis de l'identification de contenu généré par des systèmes d'intelligence artificielle;

17.6. exiger que les contenus générés par l'intelligence artificielle soient identifiés comme tels par les personnes qui les publient initialement, et que les réseaux sociaux mettent en œuvre des solutions techniques permettant aux utilisateurs et utilisatrices d'identifier facilement ces contenus, et encourager la collaboration entre les entreprises de réseaux sociaux pour

assurer l'interopérabilité des techniques de tatouage numérique pour les contenus générés par l'intelligence artificielle;

17.7. exiger que les organes de règlement extrajudiciaire des litiges, lorsqu'ils sont en place, soient indépendants et impartiaux, disposent de l'expertise nécessaire, soient facilement accessibles, fonctionnent selon des règles claires et équitables, et fassent en sorte que ces exigences soient certifiées par l'autorité de régulation nationale compétente;

17.8. promouvoir, au sein du Forum sur la gouvernance de l'internet et du Dialogue européen sur la gouvernance d'Internet, une réflexion sur la possibilité pour la communauté internet d'élaborer, dans le cadre d'un processus collaboratif et s'il y a lieu multipartite, un système d'évaluation et d'audit externe visant à déterminer si les algorithmes ne sont pas biaisés et respectent le droit à la liberté d'expression, ainsi qu'un « label de bonnes pratiques » qui pourrait être octroyé aux réseaux sociaux dont les algorithmes sont conçus pour réduire le risque de « bulles de filtrage » et de « chambres d'écho », et pour favoriser un environnement qui offre aux utilisateurs et utilisatrices une expérience pluraliste sur le plan idéologique, tout en étant sûre.

18. L'Assemblée appelle les entreprises de réseaux sociaux à éviter de prendre des mesures qui restreignent inutilement la liberté d'expression des utilisateurs et utilisatrices. Elles devraient notamment:

18.1. incorporer directement des principes juridiques fondamentaux, en particulier la liberté d'expression, dans leurs conditions générales;

18.2. faire preuve de prudence lors de la modération de contenus qui ne sont pas manifestement illégaux;

18.3. fournir aux utilisateurs et utilisatrices des conditions générales accessibles, claires et informatives sur les types de contenu autorisés sur leurs services et les conséquences en cas de non-conformité, et qui soient compréhensibles pour un large

éventail d'utilisateurs et utilisatrices, quel que soit leur niveau d'alphabétisation numérique et de maîtrise de la lecture;

18.4. informer rapidement les utilisateurs et utilisatrices que des mesures de modération ont été appliquées à leur contenu, en fournissant un exposé complet des motifs de la décision, accompagné d'une référence aux règles internes qui ont été appliquées;

18.5. s'abstenir d'appliquer le bannissement furtif aux contenus des utilisateurs et utilisatrices et informer ces personnes à chaque fois qu'une décision de déclasser ou de déréférencement a été prise;

18.6. veiller à ce que les processus automatisés de modération de contenu fassent l'objet d'une surveillance humaine et d'une évaluation rigoureuse et continue de leur fonctionnement;

18.7. mettre à disposition un système de traitement des réclamations qui soit facilement accessible, facile à utiliser et qui permette aux utilisateurs et utilisatrices de déposer une réclamation précise;

18.8. fournir aux modérateurs et modératrices humains une formation appropriée et des conditions de travail qui tiennent compte du stress psychologique important auquel ils sont soumis, et assurer une protection adéquate de leur santé;

18.9. s'abstenir de supprimer définitivement le contenu (y compris ses métadonnées) qui a été retiré conformément aux obligations légales ou aux conditions générales, en particulier lorsque le contenu en question peut servir de preuve de crimes de guerre ou d'autres crimes;

18.10. veiller à ce que les systèmes d'intelligence artificielle qu'elles développent ou utilisent respectent les normes du Conseil de l'Europe, y compris la nouvelle Convention-cadre sur l'intelligence artificielle et les droits de l'homme, la démocratie et l'État de droit; les algorithmes doivent être conçus pour respecter le droit à la liberté d'expression, et encourager la pluralité et la diversité des vues et des opinions tout en garantissant une expérience

d'utilisation sécurisée; leurs modalités de fonctionnement devraient être divulguées et les utilisateurs et utilisatrices dûment informés sur la façon dont ces algorithmes filtrent et promeuvent les contenus;

18.11. collaborer avec d'autres services en ligne afin de garantir l'interopérabilité des techniques de tatouage électronique pour les contenus générés par l'intelligence artificielle;

18.12. collaborer avec les journalistes et les organisations de vérification des faits pour lutter efficacement contre la désinformation, sur la base d'informations qui respectent les normes éthiques et professionnelles du journalisme;

18.13. promouvoir et soutenir la création d'organes indépendants de règlement extrajudiciaire des litiges, et se conformer à leurs décisions et recommandations;

18.14. soutenir le travail des organismes de contrôle tiers indépendants et se conformer à leurs décisions et recommandations;

18.15. veiller à ce que les décisions relatives à la modération de contenu soient dûment motivées et que les chercheurs aient accès à des informations complètes sur le fondement juridique et le raisonnement qui sous-tend chaque décision.

ASSEMBLEA PARLAMENTARE DEL CONSIGLIO D'EUROPA

RISOLUZIONE 2590 (2025) ⁽¹⁾

Regolamentare la moderazione dei contenuti sui *social media* per salvaguardare la libertà di espressione

ASSEMBLEA PARLAMENTARE

1. I social media sono diventati un'agorà online, dove gli utenti vanno ad esercitare il diritto alla libertà di espressione e di informazione in molti modi. Questi includono la pubblicazione di contenuti propri e la possibilità di apprezzare i contenuti pubblicati da altri, informarsi e informare gli altri e comunicare con gli altri utenti.

2. Il diritto alla libertà di espressione non è un diritto assoluto; i social media sono obbligati per legge a rimuovere contenuti illeciti quando si accorgono o vengono informati della loro esistenza sui propri servizi. Spetta inoltre ai social media combattere la diffusione di contenuti dannosi.

3. Le imprese di social media sono anche portatrici di diritti fondamentali, quali il diritto di proprietà e la libertà d'impresa, e hanno quindi voce in capitolo sul modo in cui gli utenti possono utilizzare i loro servizi e su quali contenuti possono pubblicare. Le regole di moderazione dei contenuti previste dai loro termini di servizio consentono alle imprese di social media di declassare e demonetizzare un contenuto preciso, restringere l'accesso allo stesso o

rimuoverlo a causa della sua incompatibilità con i loro termini di servizio. In casi estremi, le imprese di social media possono sospendere o addirittura chiudere l'account di un utente. I loro termini di servizio hanno carattere contrattuale e gli utenti sono tenuti a rispettarli; essi sono da 'prendere o lasciare'.

4. Le grandi imprese di social media, per lo più di proprietà di soggetti statunitensi, hanno una portata globale; le loro politiche di moderazione dei contenuti e le loro decisioni commerciali o ideologiche sui contenuti da promuovere o declassare possono esercitare una grandissima influenza sull'opinione pubblica e sulle scelte di miliardi di persone. Tuttavia, sono tenute a rispettare le leggi del Paese nel quale forniscono i propri servizi.

5. Dato il potenziale impatto che il flusso di informazioni e di comunicazioni sui social media, di fatto, ha sui comportamenti sociali e sul corretto funzionamento dei processi democratici, spetta allo Stato stabilire i principi fondamentali e il quadro istituzionale che possano correggere lo squilibrio di potere derivante da un rapporto contrattuale impari e assicurare una tutela efficace del diritto alla libertà di espressione.

6. È però essenziale che la regolamentazione pubblica della moderazione dei contenuti non abbia effetti dissuasivi sulla libertà di espressione e non sia finalizzata a

(1) Dibattito in Assemblea del 30 gennaio 2025 (8ª seduta) (cfr. Doc. 16069, relazione della Commissione Cultura, Scienza, Istruzione e Media, relatrice: on. Valentina Grippo). Testo adottato dall'Assemblea il 30 gennaio 2025 (8ª seduta).

imporre le opinioni di chi detiene il potere politico e di censurare le opinioni o le idee che possano confliggere con gli interessi acquisiti della maggioranza al potere. Inoltre, le normative nazionali non dovrebbero esagerare imponendo ai social media vincoli ingiustificati, che potrebbero tradursi in un eccesso di zelo nella rimozione dei contenuti. Tali normative e la loro attuazione devono rispettare la libertà di espressione e valutare con molta attenzione la necessità di eventuali restrizioni.

7. Il rischio determinato da politiche restrittive di moderazione dei contenuti è aumentato dalla mancanza di trasparenza nella loro attuazione. I social media sono stati accusati di una pratica chiamata « *shadow banning* » (divieto ombra) mediante la quale cancellano o declassano contenuti che affrontano questioni controverse senza informare l'utente interessato, rendendo così tali contenuti invisibili agli altri utenti. Tale pratica subdola e dissimulata dovrebbe essere vietata: priva gli utenti della possibilità di difendere efficacemente il loro diritto alla libertà di espressione.

8. La stampa e i media in generale utilizzano i social media come piattaforma per la diffusione di informazioni al pubblico. Per questo è essenziale che le pratiche di moderazione dei contenuti non incidano indebitamente sui contenuti mediatici e giornalistici che rispettano gli standard professionali e il quadro normativo nazionale.

9. La moderazione dei contenuti è svolta sempre più con mezzi automatizzati. Gli strumenti di intelligenza artificiale sono molto più efficienti dei moderatori umani nell'elaborare ad alta velocità l'immensa quantità di contenuti che circolano in rete per individuare contenuti proibiti. Tuttavia ancora oggi non hanno la capacità di comprendere del tutto le finenze dell'interazione umana (umorismo, parodia, satira, etc.) e di valutare i contenuti nel loro contesto.

10. Per tale ragione, i moderatori umani devono rimanere il fondamento di ogni sistema di moderazione dei contenuti ed essere responsabili delle decisioni da assumere nei casi in cui i sistemi automatizzati

non ne siano all'altezza. Tuttavia, la moderazione umana può non essere esente da pregiudizi e portare ad incongruenze tra un Paese e l'altro a causa di differenze culturali; è quindi essenziale stabilire norme chiare ed esaustive e garantire una formazione adeguata per fare in modo che tutti i moderatori dispongano delle conoscenze necessarie sia delle leggi in materia che delle linee guida interne dell'impresa, nonché della lingua e del contesto del Paese da cui proviene il contenuto. Tuttavia, in caso di conflitto militare tra due Paesi, i moderatori di uno dei Paesi coinvolti nel conflitto non dovrebbero moderare i contenuti provenienti dall'altro.

11. Purtroppo, malgrado il ruolo fondamentale che i moderatori umani svolgono, le loro condizioni di lavoro sono inadeguate: sono spesso sovraesposti a contenuti che possono provocare turbamenti e gravi problemi mentali e subiscono restrizioni alla libertà di parlare apertamente dei problemi che incontrano sul lavoro.

12. Gli strumenti di intelligenza artificiale generativa consentono di produrre contenuti sintetici che sono praticamente impossibili da distinguere dai contenuti generati da un essere umano. Tali contenuti possono essere molto ingannevoli ed essere uno strumento di disinformazione e manipolazione e, tra gli altri pericoli che comportano, possono incitare all'odio e alla discriminazione. È essenziale far sapere agli utenti che vi sono contenuti che appaiono autentici, ma in realtà non lo sono. A tale scopo sono particolarmente vantaggiose le tecniche di filigranatura, che tuttavia presentano numerosi inconvenienti, quali la mancanza di interoperabilità tra i servizi di social media.

13. Una valutazione indipendente dei termini di servizio, delle politiche di moderazione dei contenuti e della loro attuazione, anche al fine di individuare e promuovere le buone pratiche, potrebbe contribuire a garantire la loro conformità ai principi che difendono l'approccio alla moderazione dei contenuti fondato sui diritti umani.

14. È essenziale stabilire regole chiare e trasparenti per la risoluzione dei conflitti,

al fine di assicurare la tutela degli utenti e ridurre al minimo il rischio di essere oggetto di una decisione potenzialmente non obiettiva dell'impresa di social media, o di esser costretti a intentare un'azione in giudizio onerosa contro una società multinazionale che dispone di risorse finanziarie enormi.

15. La creazione di organismi indipendenti per la risoluzione stragiudiziale delle controversie al fine di valutare le decisioni sulla moderazione dei contenuti potrebbe rivelarsi vantaggiosa per consolidare il rispetto dei diritti fondamentali. La collaborazione tra le imprese di social media nella creazione di tali organismi potrebbe auspicabilmente anche facilitare la composizione delle controversie.

16. Come indicato dall'Assemblea nella propria Risoluzione 2281 (2019) dal titolo « Social media: interazioni sociali o minacce per i diritti umani? », le imprese di social media dovrebbero utilizzare algoritmi che favoriscano la diversità delle fonti, degli argomenti e delle opinioni e garantiscano la qualità delle informazioni disponibili, riducendo così il rischio delle « bolle di filtraggio » e delle « casse di risonanza ».

17. Alla luce di tali considerazioni, l'Assemblea invita gli Stati membri del Consiglio d'Europa a rivedere la propria legislazione al fine di tutelare meglio il diritto alla libertà di espressione sui social media. A tale proposito, dovrebbero in particolare:

17.1. esigere che i social media rispettino i diritti fondamentali degli utenti, ivi compresa la libertà di espressione, nelle loro politiche di moderazione dei contenuti e nelle loro pratiche di attuazione;

17.2. esigere che le piattaforme di social media giustificino qualunque misura assunta per moderare i contenuti forniti dalla stampa o dai fornitori di servizi di media prima della loro attuazione, e dar loro la possibilità di rispondere entro un lasso di tempo adeguato;

17.3. in collaborazione con la stampa o le organizzazioni dei media, realizzare un sistema di verifica degli account dei media e dei giornalisti, insieme a meccanismi solidi in grado di proteggerli da molestie

online, attacchi informatici e frodi, e sviluppare linee guida sui social media per la stampa o le organizzazioni dei media relative alla pubblicazione di informazioni su questioni sensibili, al fine di evitare inutili restrizioni di moderazione su questo tipo di contenuti;

17.4. prevedere norme minime che definiscano le condizioni di lavoro dei moderatori umani, ivi compreso l'obbligo di un'adeguata formazione per lo svolgimento delle loro mansioni spesso stressanti e dell'accesso a un'adeguata assistenza psicologica e a servizi di salute mentale in caso di necessità;

17.5. firmare e ratificare la Convenzione Quadro del Consiglio d'Europa sull'Intelligenza Artificiale, i Diritti dell'Uomo, la Democrazia e lo Stato di Diritto (STCE N. 225, « Convenzione di Vilnius ») e adottare o mantenere misure per garantire che siano previsti requisiti adeguati di trasparenza e controllo adattati ai contesti e ai rischi specifici, per affrontare le sfide dell'individuazione di contenuti generati dai sistemi di intelligenza artificiale;

17.6. esigere che i contenuti generati dall'intelligenza artificiale siano divulgati come tali dalle persone che li pubblicano inizialmente e che i social media attuino soluzioni tecniche che consentano agli utenti di individuare facilmente tali contenuti, e incoraggiare la collaborazione tra le imprese di social media per assicurare l'interoperabilità delle tecniche di filigranatura per i contenuti generati dall'intelligenza artificiale;

17.7. esigere che gli organismi indipendenti per la risoluzione stragiudiziale delle controversie, una volta istituiti, siano indipendenti e imparziali, abbiano le competenze necessarie, siano facilmente accessibili, e operino secondo regole chiare ed eque, e provvedano alla certificazione di tali requisiti da parte dell'autorità di regolamentazione nazionale competente;

17.8. promuovere, in seno al Forum sulla Governance di Internet e nel Dialogo Europeo sulla Governance di Internet, una riflessione sulla possibilità che la comunità

di internet elabori, attraverso un approccio collaborativo e, ove opportuno, un processo che coinvolga una pluralità di soggetti interessati, un sistema esterno di valutazione e verifica finalizzato a determinare se gli algoritmi sono imparziali e rispettano il diritto alla libertà di espressione, e un « sigillo di buone pratiche » da assegnare ai social media i cui algoritmi siano concepiti in modo da ridurre il rischio delle « bolle di filtraggio » e delle « casse di risonanza », e promuovere un ambiente che offra agli utenti un'esperienza ideologicamente trasversale, ma sicura.

18. L'Assemblea invita le imprese di social media a evitare di adottare misure che limitino inutilmente la libertà di espressione degli utenti. In particolare, esse dovrebbero:

18.1. integrare direttamente nei propri termini di servizio i principi dei diritti fondamentali, e in particolare la libertà di espressione;

18.2. procedere con cautela nel moderare contenuti che non siano evidentemente illeciti;

18.3. fornire agli utenti termini di servizio che siano facilmente accessibili, chiari e informativi sui tipi di contenuti ammessi sui loro servizi e sulle conseguenze del mancato rispetto dei termini di servizio, e che risultino comprensibili a un'ampia gamma di utenti a prescindere dai diversi livelli di alfabetizzazione digitale e competenza nella lettura;

18.4. informare rapidamente gli utenti di eventuali interventi di moderazione effettuati sui loro contenuti, fornendo un resoconto completo dei motivi per cui si è deciso di intervenire, accompagnati da un rinvio alle regole interne che sono state applicate;

18.5. astenersi dalla pratica di *shadow banning* dei contenuti degli utenti e informarli di ogni caso di declassamento o cancellazione;

18.6. assicurare che i processi automatizzati di moderazione dei contenuti siano

sottoposti al controllo umano e a una valutazione rigorosa e continua del loro funzionamento;

18.7. mettere a disposizione un sistema per la gestione dei reclami che sia facilmente accessibile, facile da utilizzare, e consenta agli utenti di presentare reclami precisi;

18.8. prevedere per i moderatori umani una formazione adeguata e condizioni di lavoro che prestino attenzione al grave stress psicologico al quale sono sottoposti, e garantire un'adeguata tutela della loro salute;

18.9. astenersi dall'eliminare definitivamente contenuti (ivi inclusi i relativi metadati) che siano stati rimossi conformemente agli obblighi di legge o ai termini di servizio, in particolare qualora i contenuti in questione possano fungere da prova di crimini di guerra o altri reati;

18.10. garantire che i sistemi di intelligenza artificiale che essi sviluppano o utilizzano rispettino le norme del Consiglio d'Europa, ivi compresa la nuova Convenzione Quadro sull'Intelligenza Artificiale, i Diritti dell'Uomo, la Democrazia e lo Stato di Diritto; gli algoritmi dovrebbero essere concepiti per rispettare il diritto alla libertà di espressione, e incoraggiare la pluralità e la diversità dei punti di vista e delle opinioni pur garantendo all'utente un'esperienza sicura; le loro modalità operative dovrebbero essere divulgate e gli utenti dovrebbero essere debitamente informati sul modo in cui detti algoritmi filtrano e promuovono i contenuti;

18.11. collaborare con altri servizi online per garantire l'interoperabilità delle tecniche di filigranatura per i contenuti generati dall'intelligenza artificiale;

18.12. collaborare con giornalisti e organizzazioni che si occupano di *fact-checking* per combattere efficacemente la disinformazione con informazioni che rispettano gli standard etici e professionali del giornalismo;

18.13. promuovere e sostenere la creazione di organismi indipendenti per la risoluzione stragiudiziale delle controver-

sie, e rispettare le loro decisioni e raccomandazioni;

18.14. sostenere il lavoro degli organismi di controllo terzi e rispettare le loro decisioni e raccomandazioni;

18.15. assicurare che le decisioni relative alla moderazione dei contenuti siano debitamente motivate e che i ricercatori abbiano accesso a informazioni complete riguardanti il fondamento giuridico e il ragionamento alla base di ogni decisione.



190122141910