



Doc. 16089

08 janvier 2025

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

Rapport¹

Commission de la culture, de la science, de l'éducation et des médias

Rapporteure: Mme Valentina GRIPPO, Italie, Alliance des démocrates et des libéraux pour l'Europe

Résumé

Les réseaux sociaux permettent à chaque utilisateur et utilisatrice de partager des contenus avec un large public et de participer à des communautés et des réseaux virtuels. Ces activités sont réglementées contractuellement par les règles de modération de contenu qui sont énoncées dans les conditions générales des entreprises de réseaux sociaux. Les utilisateurs et utilisatrices sont tenus de respecter ces règles, qui sont «à prendre ou à laisser».

Les États membres doivent mettre en place les principes de base et le cadre institutionnel qui peuvent corriger le déséquilibre des pouvoirs qui découle de cette relation contractuelle inégale et assurer la protection effective du droit à la liberté d'expression en ligne. Il est cependant essentiel que la réglementation publique de la modération de contenu ne conduise pas les entreprises de réseaux sociaux à un excès de zèle dans la suppression de contenus.

Le présent rapport appelle les entreprises de réseaux sociaux à s'abstenir de mettre en œuvre des politiques qui limitent indûment la liberté d'expression des utilisateurs et utilisatrices. Les conditions générales qu'elles appliquent doivent être claires, facilement accessibles et fondées sur les principes des droits fondamentaux. Ces entreprises doivent fournir aux modérateurs et modératrices humains une formation complète ainsi que des conditions de travail appropriées (y compris des soins de santé mentale), et utiliser de manière efficace des outils de modération de contenu automatisés qui font l'objet d'une surveillance humaine et d'une évaluation rigoureuse et continue. Elles doivent informer les utilisateurs et utilisatrices lorsque des mesures de modération de contenu sont prises, et ce rapidement et de manière circonstanciée. Leurs systèmes de traitement des réclamations doivent être accessibles et faciles à utiliser. Enfin, elles doivent contribuer à la mise en place d'organismes indépendants de résolution extrajudiciaire des litiges.

1. Renvoi en commission: [Doc. 15555](#), Renvoi 4705 du 23 janvier 2023.



Sommaire

Page

A. Projet de résolution	3
B. Exposé des motifs par Mme Valentina Grippo, rapporteure	7
1. Introduction	7
2. La réglementation de l'Union européenne en matière de modération de contenu	7
2.1. Le Règlement sur les services numériques	8
2.2. Le règlement européen sur la liberté des médias	9
3. Les activités normatives du Conseil de l'Europe	10
4. La modération de contenu dans la pratique	12
4.1. Comparaison des pratiques	12
4.2. Sujets de préoccupation	12
5. Options de règlement des conflits	14
5.1. Système interne de traitement des réclamations	14
5.2. Règlement extrajudiciaire des litiges	14
5.3. Recours judiciaire	15
6. Trouver le bon équilibre des droits	16
7. Améliorer la modération de contenu	18
7.1. Intégrer les principes des droits fondamentaux dans les conditions générales	18
7.2. Clarté des conditions générales	19
7.3. Formation adéquate et bonnes conditions de travail pour les modérateurs et modératrices	19
7.4. Utiliser l'intelligence artificielle de manière efficace	20
7.5. Le rôle des systèmes de recommandation	22
7.6. Résumé	23

A. Projet de résolution²

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 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

2. Projet de résolution adopté à l'unanimité par la commission le 4 décembre 2024.

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.

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 inconvénients, 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'intenter 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. 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;

17.4. 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;

17.5. 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;

17.6. 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;

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

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

18.14. 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.

B. Exposé des motifs par M^{me} Valentina Grippo, rapporteure

1. Introduction

1. Les réseaux sociaux³ ont le pouvoir, et dans certains cas l'obligation légale, de filtrer, d'empêcher la diffusion, de déclasser, de bloquer et, le cas échéant, de retirer les contenus illégaux ou potentiellement nuisibles. En outre, leurs conditions générales d'utilisation établissent des règles sur la nature du contenu et le mode de comportement qui sont acceptables sur leurs services et la manière dont ils peuvent en restreindre la fourniture. Il leur arrive cependant d'adopter des mesures relatives à la mise en œuvre de leurs politiques de modération de contenu qui sont contraires au droit à la liberté d'expression.

2. Si les entreprises de réseaux sociaux sont elles-mêmes porteuses de droits fondamentaux tels que le droit de propriété et la liberté d'entreprise, il incombe à la réglementation publique de fixer les principes fondamentaux et le cadre institutionnel nécessaires pour assurer la protection effective des droits fondamentaux des utilisateurs et des utilisatrices, et en particulier du droit à la liberté d'expression.

3. L'objectif de mon rapport est d'examiner la manière de protéger le droit à la liberté d'expression sur les réseaux sociaux en veillant à ce que le droit de propriété et la liberté d'entreprendre de ces entreprises ne soient pas injustement limités.

4. A cette fin j'examinerai les questions suivantes:

- les obligations que les entreprises de réseaux sociaux doivent respecter au regard du droit de l'Union européenne, notamment en ce qui concerne le contenu, l'application et l'exécution de leurs conditions générales, en particulier les mesures de modération de contenu et les mécanismes de contrôle et garantissant le respect de ces obligations;
- les mesures spécifiques préconisées par le droit souple du Conseil de l'Europe;
- les politiques de modération de contenu des grandes entreprises de réseaux sociaux et les difficultés éventuelles qu'elles rencontrent pour intégrer la législation européenne et les normes du Conseil de l'Europe;
- les normes et les mesures qu'il faudrait introduire pour tenir compte des évolutions récentes et futures.

5. Mon analyse, qui s'appuie sur les contributions de plusieurs expert·es que nous avons entendus⁴, bénéficie des informations que j'ai recueillies lors de ma visite d'information à Dublin en juillet 2024, au cours de laquelle j'ai rencontré des représentant·es de trois grandes entreprises technologiques: Meta, TikTok et Google. J'ai également tenu des réunions fructueuses avec les autorités irlandaises de régulation des médias et de protection des données.

2. La réglementation de l'Union européenne en matière de modération de contenu

6. Depuis 2000, les services en ligne sont réglementés au niveau de l'Union européenne par la [Directive sur le commerce électronique](#). Cette directive prévoit une dérogation en matière de responsabilité pour les fournisseurs d'hébergement lorsqu'ils n'ont pas effectivement connaissance des activités ou des contenus illicites ou lorsqu'ils agissent promptement dès le moment où ils en prennent connaissance. Les États membres, cependant, ne peuvent pas imposer aux prestataires de services de la société de l'information une obligation générale de surveiller les informations qu'ils transmettent ou stockent, ou une obligation générale de rechercher activement des faits ou des circonstances révélant des activités illicites.

3. Aux fins du présent rapport, le terme «réseaux sociaux» fait référence aux services en ligne qui permettent aux utilisateurs et utilisatrices de partager du contenu avec un vaste public et de participer à des communautés et des réseaux en ligne. Cette définition englobe des plateformes telles que Facebook, Tik Tok, YouTube et d'autres services similaires. Malgré cette définition de travail, les termes «plateformes» et «plateformes en ligne» sont utilisés dans le texte afin d'être conformes à la terminologie utilisée dans d'autres contextes, en particulier dans la législation de l'Union européenne.

4. Je tiens à remercier en particulier: M. Lubos Kuklis, équipe Digital Services Act (DSA), Commission européenne, Bruxelles; M. Jack Goodman, BBC, Royaume-Uni; M. Christian Hannibal, chef des politiques publiques, Tik Tok, Danemark; M. Mark David Cole, professeur de droit des médias et des télécommunications à l'Université du Luxembourg; et M^{me} Gemma Shields, responsable de la politique de sécurité en ligne (droits de l'homme et transparence) à l'Ofcom, Royaume-Uni.

7. Ces règles générales ont été complétées par l'Union européenne au cours de la période 2000-2020 par une série de mesures législatives et de corégulation spécifiques couvrant les [services de médias audiovisuels](#), le [droit d'auteur](#), le [terrorisme](#), les [abus sexuels commis sur des enfants](#), le [discours de haine](#) et la [désinformation](#).

8. En 2019, la Commission européenne a lancé le processus d'adoption d'un paquet réglementaire complet en matière de services numériques, le [paquet «Législation sur les services numériques»](#). À l'issue de ce processus, deux nouveaux règlements, à savoir le Règlement sur les services numériques et le Règlement sur les marchés numériques, ont été promulgués en 2022.

2.1. Le Règlement sur les services numériques

9. Le [Règlement sur les services numériques \(DSA\)](#), pour Digital Services Act) vise à prévenir les activités illicites et préjudiciables en ligne et la propagation de la désinformation sur les services intermédiaires. Le DSA met l'accent sur les «très grandes plateformes en ligne» et les «très grands moteurs de recherche en ligne», qui, en raison de leur taille au regard des destinataires de ces services⁵, sont soumis à une réglementation spéciale.

10. L'article 14 du DSA réglemente les activités des services intermédiaires (dont les réseaux sociaux) qui visent à détecter, à identifier et à traiter les contenus fournis par les utilisateurs et les utilisatrices qui sont incompatibles avec leurs conditions générales.

11. En principe, le DSA défend la liberté contractuelle des fournisseurs de services intermédiaires, mais fixe des règles concernant le contenu, l'application et le respect de leurs conditions générales «dans un souci de transparence, de protection des destinataires du service et de prévention de conséquences inéquitables ou arbitraires» (considérant 45 du DSA). Les conditions générales des fournisseurs devraient clairement indiquer les motifs au titre desquels ils peuvent restreindre la fourniture de leurs services aux utilisateurs et utilisatrices.

12. Les services intermédiaires doivent expliquer, dans leurs conditions générales, les restrictions qu'ils imposent en ce qui concerne le contenu publié par leurs utilisateurs et utilisatrices et la manière dont ils modèrent ce contenu. Ils doivent en particulier inclure des renseignements concernant les politiques, les procédures, les mesures et les outils utilisés à des fins de modération des contenus, y compris la prise de décision fondée sur des algorithmes et le réexamen par un être humain, ainsi que le règlement intérieur de leur système interne de traitement des réclamations. Ces renseignements doivent être énoncés dans un langage clair, simple, intelligible, aisément abordable et dépourvu d'ambiguïté, et doivent être mis à la disposition du public dans un format facilement accessible et lisible par machine.

13. Les services intermédiaires doivent appliquer et faire respecter ces règles de manière diligente, objective et proportionnée en tenant dûment compte des droits et des intérêts légitimes de toutes les parties impliquées, et notamment des droits fondamentaux des utilisateurs et utilisatrices, tels que la liberté d'expression, la liberté et le pluralisme des médias, et d'autres libertés et droits fondamentaux tels qu'ils sont consacrés par la [Charte des droits fondamentaux de l'Union européenne](#).

14. De plus, les services intermédiaires doivent informer les utilisateurs et utilisatrices de toute modification importante des conditions générales.

15. Dans le cas des services intermédiaires qui s'adressent principalement à des mineurs ou qui sont utilisés majoritairement par ces derniers, les conditions générales doivent être formulées de manière à être facilement compréhensibles pour ce public.

16. En ce qui concerne les systèmes de recommandation, les conditions générales doivent exposer dans un langage clair et intelligible leurs principaux paramètres, y compris la manière dont les informations suggérées à l'utilisateur sont déterminées, et les raisons de l'importance relative de ces paramètres. Les conditions générales doivent également décrire les options dont disposent les utilisateurs pour modifier ou influencer ces principaux paramètres. Lorsque plusieurs options sont disponibles pour déterminer l'ordre relatif des informations présentées aux utilisateurs, le service doit également permettre à l'utilisateur de sélectionner et de modifier à tout moment son option préférée de manière simple (article 27 du DSA).

5. Les très grandes plateformes en ligne et les très grands moteurs de recherche en ligne ont un nombre mensuel moyen de destinataires actifs du service dans l'Union européenne égal ou supérieur à 45 millions et sont désignés comme tels par la Commission européenne (article 33, paragraphe 1, du DSA).

17. Les décisions de modération de contenu doivent être notifiées aux utilisateurs avec un exposé des motifs (article 17 du DSA). Par conséquent, des systèmes de traitement des réclamations effectifs doivent être en place (article 20 du DSA), et les utilisateurs et utilisatrices doivent pouvoir choisir tout organe de règlement extrajudiciaire des litiges qui a été certifié par le coordinateur pour les services numériques compétent (article 21 du DSA).

18. Les États membres de l'Union européenne devraient respecter les droits fondamentaux à un recours effectif et à accéder à un tribunal impartial, comme le prévoit l'article 47 de la [Charte des droits fondamentaux](#) et les articles 6 et 13 de la [Convention européenne des droits de l'homme](#) (STE n° 5, «la Convention»)

19. Les très grandes plateformes en ligne et les très grands moteurs de recherche en ligne doivent évaluer les risques systémiques découlant de la conception, du fonctionnement et de l'utilisation de leurs services, ainsi que des abus potentiels par les utilisateurs et utilisatrices, et devraient prendre des mesures d'atténuation appropriées, dans le respect des droits fondamentaux (article 34 du DSA).

20. Lorsqu'ils procèdent à cette évaluation des risques, les très grandes plateformes en ligne et les très grands moteurs de recherche en ligne doivent, entre autres, prendre en considération l'influence, entre autres, de leurs systèmes de modération de contenu, les conditions générales applicables et leur mise en œuvre.

21. Les très grandes plateformes en ligne et les très grands moteurs de recherche en ligne doivent mettre en place des mesures d'atténuation raisonnables, proportionnées et effectives, adaptées aux risques systémiques spécifiques mentionnés plus haut, en accordant une attention particulière aux effets de ces mesures sur les droits fondamentaux. Ces mesures peuvent notamment inclure: l'adaptation de leurs conditions générales et de la mise en application de celles-ci, ainsi que des processus de modération des contenus (article 35 du DSA).

22. Les services intermédiaires doivent garantir un niveau adéquat de transparence et de responsabilisation (considérant 49 du DSA). À cette fin, ils doivent mettre à la disposition du public des rapports clairs et facilement compréhensibles sur leurs activités de modération de contenu. Ces rapports doivent être publiés au moins une fois par an, dans un format lisible par machine et d'une manière facilement accessible (article 15 du DSA). Des obligations plus strictes en matière de rapports incombent aux fournisseurs de plateformes en ligne (article 24 du DSA) et en particulier aux très grandes plateformes en ligne et aux très grands moteurs de recherche en ligne (article 42 du DSA), qui font l'objet d'audits indépendants (article 37 du DSA).

23. La Commission européenne gère une base de données consacrée à la transparence dans le cadre du DSA ([DSA Transparency Database](#)). La base de données contient les décisions et les exposés des motifs des plateformes en ligne lorsque ces dernières retirent des informations ou limitent d'une autre manière la disponibilité d'informations et l'accès à des informations. Elle vise à garantir la transparence et à permettre le contrôle des décisions relatives à la modération de contenu des plateformes en ligne ainsi que le suivi de la diffusion de contenus illicites en ligne (considérant 66 du DSA).

2.2. Le règlement européen sur la liberté des médias

24. Le [règlement européen sur la liberté des médias \(EMFA\)](#), pour European Media Freedom Act) réglemente le pluralisme et l'indépendance des médias dans l'Union européenne.

25. En ce qui concerne la modération de contenu, l'article 18 du EMFA protège les fournisseurs de services de médias (par exemple les chaînes de télévision) contre le retrait injustifié par les très grandes plateformes en ligne de contenus médiatiques considérés comme incompatibles avec leurs conditions générales. Les très grandes plateformes en ligne devront justifier toute mesure de modération de contenu avant sa mise en œuvre et donner au fournisseur de services de médias la possibilité de répondre dans un délai de 24 heures ou dans un délai plus court dans les situations de crise prévues par le DSA. Cette procédure d'avertissement rapide ne s'applique pas, en revanche, lorsque des décisions relatives à la modération de contenu sont prises conformément aux règles du DSA et de la Directive «Services de médias audiovisuels», notamment en ce qui concerne les obligations de retirer des contenus illicites, de protéger les mineurs et d'atténuer les risques systémiques.

3. Les activités normatives du Conseil de l'Europe

26. L'Assemblée parlementaire et en particulier la commission de la culture, de la science, de l'éducation et des médias traitent depuis plusieurs années la question de la réglementation de la liberté d'expression sur l'internet. Déjà en 2012, dans le rapport intitulé «[La protection de la liberté d'expression et d'information sur l'internet et les médias en ligne](#)»⁶, nous avons exprimé la crainte que des opérateurs privés ayant une position dominante sur le marché des services internet ne restreignent de manière abusive l'accès aux informations et leur diffusion, sans en informer leurs usagers et usagères et en violation de leurs droits.

27. Dans la [Résolution 1877 \(2012\)](#), l'Assemblée a donc appelé les États membres à veiller à ce que les intermédiaires de l'internet soient transparents (paragraphe 11.3) et tenus pour responsables des violations du droit de leurs utilisateurs et utilisatrices à la liberté d'expression et d'information (paragraphe 11.6).

28. Par la suite, d'autres rapports de la commission de la culture, de la science, de l'éducation et des médias ont abordé la question de la réglementation publique de la liberté d'expression sur internet sous des angles différents et l'Assemblée a élaboré des lignes directrices visant tant l'élaboration de la législation nationale que les normes d'autorégulation des opérateurs privés⁷.

29. La modération de contenu et les effets des algorithmes utilisés par les plateformes internet sur les droits humains ont également fait l'objet d'études détaillées du secteur intergouvernemental du Conseil de l'Europe. Les normes visant à clarifier les responsabilités des autorités publiques et à encadrer l'action des intermédiaires d'Internet sont énoncées dans plusieurs recommandations du Comité des Ministres aux États membres la [Recommandation CM/Rec\(2018\)2 sur les rôles et les responsabilités des intermédiaires d'internet](#), la [Recommandation CM/Rec\(2022\)13 sur les effets des technologies numériques sur la liberté d'expression](#), la [Recommandation CM/Rec\(2022\)16 sur la lutte contre le discours de haine](#) et la [Recommandation CM/Rec\(2020\)1 sur les impacts des systèmes algorithmiques sur les droits de l'homme](#). De plus, la [Note d'orientation sur la modération de contenu \(2021\)](#) et la [Note d'orientation sur la hiérarchisation des contenus d'intérêt public en ligne \(2021\)](#) prévoient des principes et des lignes d'action visant à maintenir l'accès du public à des informations de qualité. En outre, la [Note d'orientation sur la lutte contre la propagation de la désinformation et de la désinformation en ligne par le biais de la vérification des faits et de la conception de plateformes dans le respect des droits de l'homme \(2023\)](#) se concentre sur quatre domaines d'action: la vérification des faits, les solutions de conception de plateformes, l'autonomisation des utilisateurs et utilisatrices, et l'éducation aux médias.

30. La [Recommandation CM/Rec\(2018\)2](#) affirme que «les États ont l'obligation fondamentale de protéger les droits de l'homme et les libertés fondamentales dans l'environnement numérique. Tous les cadres réglementaires, y compris les modes d'autorégulation ou de corégulation, devraient prévoir des mécanismes de surveillance efficaces pour être conformes à cette obligation et devraient être assortis de possibilités de recours appropriées» (ligne directrice 1.1.3.).

31. Le document normatif le plus approfondi du Conseil de l'Europe en la matière est la [Note d'orientation sur la modération de contenu \(2021\)](#). Elle fournit des conseils pratiques aux États membres du Conseil de l'Europe, qui tiennent compte des bonnes pratiques existantes pour l'élaboration de politiques, la régulation et le recours à la modération de contenu en ligne conformément à leurs obligations en vertu de la Convention européenne des droits de l'homme. Elle s'adresse également aux intermédiaires d'internet qui ont leurs propres responsabilités en matière de droits humains.

32. La note d'orientation énumère les principes fondamentaux d'une approche de la modération de contenu fondée sur les droits humains:

- a. Transparence: la transparence est essentielle pour garantir la responsabilité, la flexibilité, la non-discrimination, l'efficacité et la proportionnalité, ainsi que pour l'identification et l'atténuation des conflits d'intérêts. Des normes minimales devraient être définies pour évaluer si la modération de contenu en question atteint ses objectifs spécifiques et un examen indépendant d'au moins un échantillon représentatif de cas de modération de contenu devrait être mené.
- b. Les droits de l'homme par défaut: en vertu de la Convention européenne des droits de l'homme, les droits humains sont la norme et les restrictions sont l'exception. Cette approche doit guider l'élaboration des politiques en matière de modération de contenu. Il est nécessaire d'identifier de manière proactive les droits qui pourraient être menacés avant de lancer tout processus de modération de contenu. De plus, le droit à un recours effectif (article 13 de la Convention) doit être respecté. Enfin, il ne suffit pas

6. [Doc. 12874](#).

7. [AS/Cult/Inf\(2024\)12](#).

de procéder à un examen préalable des mesures d'autorégulation ou de corégulation pour garantir le respect des droits humains; il est également indispensable d'effectuer un examen fréquent de l'effet ou des effets de ces mesures.

- c. Identification des problèmes et des cibles: les interventions politiques ayant pour but de réduire les risques au minimum doivent être clairement reconnues en tant que telles, afin d'atténuer les problèmes particuliers de cette approche, l'État assumant sa part de responsabilité. Elles devraient également être assorties d'objectifs clairs, de mécanismes d'ajustement et de supervision, d'une réelle protection de la liberté d'expression, ainsi que d'outils permettant de repérer les effets contre-productifs. Lorsque la modération de contenu est effectuée dans le cadre d'un système d'autorégulation ou de corégulation, il faudrait également prévoir des mécanismes permettant de reconcevoir, d'adapter ou d'abandonner le système, si les normes minimales ne sont pas respectées ou si la nature du problème évolue de telle sorte que l'approche retenue n'est pas efficace.
- d. Décentralisation significative: une modération décentralisée, multipartite, rémunérée, habilitée et indépendante est essentielle pour tenir compte des particularités régionales lorsqu'il s'agit de traiter les types de contenu les plus difficiles. Les plateformes en ligne devraient mettre en place des conseils multipartites pour les aider à évaluer les problèmes les plus difficiles, pour examiner des questions émergentes et pour s'opposer aux plus hauts niveaux de direction de l'entreprise. En outre, des données adéquates doivent être mises à la disposition de la société civile et des chercheurs techniques et universitaires pour faciliter une analyse continue.
- e. Communication avec l'utilisateur – les restrictions aux libertés fondamentales doivent respecter les normes en matière de droits humains et être les plus transparentes possibles. Les conditions générales devraient être formulées d'une manière claire et accessible. L'application de ces règles devrait également être prévisible, conformément aux principes fondamentaux des droits humains. De plus, les utilisateurs et utilisatrices doivent pouvoir déposer un recours auprès de l'entreprise de la manière la plus spécifique possible. Si le retrait du contenu n'est pas urgent, les utilisateurs et utilisatrices concernés devraient recevoir des informations claires sur les raisons pour lesquelles ce contenu peut avoir enfreint les conditions générales ou la loi, avoir le droit de défendre leur cause dans un délai déterminé et, dans tous les cas, avoir le droit à un véritable recours. Les contenus qui doivent être mis hors ligne le plus vite possible doivent être bien définis et le processus de révision, de suppression et, si nécessaire, de remise en ligne doit être prévisible, responsable et proportionné.
- f. Des garanties administratives de haut niveau: un cadre juridique clair et prévisible est indispensable pour veiller à ce que les restrictions soient prévues par la loi. Cela nécessite également une transparence réelle sur la gouvernance, les processus décisionnels et les détails englobant le retrait ou non d'un contenu (comment? quand? pourquoi? combien de fois? quel contenu?) et pour quelles raisons. Des rapports efficaces sur la transparence permettront aux entreprises et au grand public d'identifier le jeu⁸ des mécanismes de plainte des entreprises. En outre, la publication de données suffisantes sur les décisions prises et la mise à disposition, à un organisme indépendant et impartial, d'un nombre suffisant d'échantillons, en vue d'un contrôle proactif, sont nécessaires pour assurer la cohérence et l'indépendance des mécanismes de contrôle. Les intermédiaires d'internet doivent également tenir compte de ces constatations de manière significative. En outre, il convient également d'accorder une attention appropriée aux droits du travail et à la santé mentale de l'ensemble des travailleurs et travailleuses chargés de modérer des contenus susceptibles d'être choquants ou perturbants ou d'avoir des répercussions psychologiques sur ces personnes. La protection de la vie privée et des données doit être garantie. Les États devraient prendre en considération les droits des victimes de contenus illicites afin d'assurer un soutien total visant à annuler ou à atténuer les dommages causés. Des mesures appropriées sont également nécessaires pour indemniser les victimes de retraits injustifiés et éviter que de tels problèmes ne se produisent.
- g. Traiter les particularités de l'autorégulation et de la corégulation en matière de modération de contenu: l'ampleur du flux de communication avec les utilisateurs et utilisatrices diffère considérablement lorsqu'il s'agit de médias en ligne. Par conséquent, les hypothèses fondées sur l'expérience de l'autorégulation et de la corégulation des médias traditionnels peuvent être trompeuses dans le contexte de l'autorégulation de la plupart des intermédiaires d'internet. Il est essentiel que le rôle de l'État soit précisément défini afin que l'obligation de rendre des comptes puisse prendre effet.

8. Il s'agit de «la manipulation délibérée, pas nécessairement illégale, ou [de] l'abus des systèmes de plainte des intermédiaires d'internet». Voir page 25 de la Note d'orientation.

4. La modération de contenu dans la pratique

4.1. Comparaison des pratiques

33. La modération de contenu est un terme qui englobe toutes les activités entreprises par les services intermédiaires dans le but de détecter, d'identifier et de traiter les contenus créés par les utilisateurs et utilisatrices qui sont soit illégaux, soit incompatibles avec les conditions générales. Dans la pratique, cet objectif est mis en œuvre par des mesures qui visent le contenu (déclassement, démonétisation, désactivation de l'accès, suppression) et des mesures qui touchent l'utilisateur (résiliation ou suspension du compte d'un utilisateur ou d'une utilisatrice).⁹

34. Chaque réseau social possède des politiques de modération de contenu, des règles d'application et des pratiques pour les faire respecter qui sont spécifiques, ce qui rend toute analyse comparative difficile. En outre, il convient de faire preuve de prudence lors de l'examen des rapports de transparence de ces réseaux sociaux, car il existe plusieurs différences méthodologiques, les périodes d'observation ne correspondent pas toujours aux six mois exigés par l'Union européenne et un nombre important d'informations font défaut¹⁰.

35. En 2024, une étude a été publiée sur [les médias et la société après la rupture technologique](#), qui examine les politiques de modération de contenu de Facebook, YouTube, TikTok, Reddit et Zoom, ainsi que leur mise en œuvre concrète. Elle passe en revue les politiques en matière de contenu, les règles d'application et leur mise en œuvre, ainsi que la transparence.

36. Les trois plus grands réseaux sociaux (Facebook, YouTube et TikTok), qui font face à des défis similaires compte tenu de leur taille, ont des politiques semblables en matière de modération de contenu:

- des politiques couvrant l'ensemble de la plateforme contre un grand nombre de contenus similaires;
- une mise en œuvre à plusieurs niveaux, qui consiste à interdire certains types de contenus et à limiter l'accès à d'autres types de contenus ou la diffusion de ces contenus;
- la mise en garde des utilisateurs et utilisatrices qui publient des contenus contraires aux politiques et la suppression de leur compte en cas de violations répétées.

37. L'approche de Reddit par canal de discussion est très différente, car la plupart des règles sont fixées par les utilisateurs et utilisatrices eux-mêmes en fonction des types de discussions qu'ils et elles souhaitent avoir au sein de groupes spécifiques.

38. Les réseaux sociaux appliquent généralement leurs politiques de modération de contenu de manière proactive en recherchant les contenus qui violent les politiques, mais certains le font de manière réactive, en réponse aux rapports des utilisateurs et utilisatrices.

39. En ce qui concerne les conséquences sur les utilisateurs et utilisatrices qui enfreignent les conditions générales, la plupart des réseaux sociaux utilisent des systèmes d'avertissement qui ne sont pas assez clairs pour l'internaute, sachant que YouTube est une exception notable positive à cet égard. Ce manque de clarté, qui permet aux réseaux sociaux d'ajuster leurs politiques en réaction à des événements sans avoir à communiquer chaque changement au public, réduit néanmoins la transparence et l'obligation de rendre des comptes.

40. S'il existe des similitudes dans l'application des politiques qui concernent les contenus illégaux, dont le cadre est fixé par des obligations légales, les mécanismes d'application peuvent néanmoins différer. Chacun des réseaux sociaux examinés ici indique clairement dans ses conditions générales que les utilisateurs et utilisatrices qui publient des contenus illégaux perdront leur compte immédiatement, sans avis préalable ni avertissement.

4.2. Sujets de préoccupation

41. La modération de contenu crée un conflit entre les droits fondamentaux des utilisateurs et utilisatrices, tels que la liberté d'expression, et le droit de propriété et la liberté d'entreprendre des entreprises de réseaux sociaux. En principe, les réseaux sociaux ont le droit de décider quel contenu est acceptable sur leurs

9. Voir article 3, alinéa t) du DSA.

10. Voir l'article «[Content moderation: Key facts to learn from Facebook, Instagram, X and TikTok transparency reports](#)». (Modération en ligne: ce que nous apprennent les rapports de transparence de Facebook, Instagram, TikTok).

services, et comment les utilisateurs et utilisatrices peuvent les utiliser. Ces règles sont contenues dans leurs conditions générales, qui ont un caractère contractuel, et les utilisateurs et utilisatrices sont tenus de respecter ces règles, qui sont «à prendre ou à laisser», sans possibilité de négocier aucune des clauses.

42. Cette réglementation du contenu par contrat soulève de sérieuses préoccupations en termes de protection de la liberté d'expression pour les raisons suivantes¹¹:

- des normes moins strictes en matière de liberté d'expression: si les réseaux sociaux sont en principe libres de restreindre le contenu en se fondant sur la liberté contractuelle, certaines politiques, telles que la suppression de contenu «pour quelque raison que ce soit ou sans raison», ne sont pas conformes à l'obligation de respecter les droits fondamentaux. En outre, le fait que les règles qui régissent la modération de contenu ne soient pas guidées par les principes de nécessité et de proportionnalité pose problème.
- le manque de transparence et d'obligation de rendre des comptes: le manque de transparence concernant la mise en œuvre des conditions générales a une incidence sur la capacité à tenir les entreprises responsables des déclassements de contenu injustifiés, arbitraires ou discriminatoires. Il peut s'agir d'un manque de clarté du champ d'application d'une règle liée au contenu, d'un manque de disponibilité des conditions générales dans certaines langues ou d'un manque de granularité dans les rapports de modération de contenu.
- l'absence de garanties procédurales et de recours: il existe de nombreux exemples à ce sujet, notamment le manque de clarté concernant le moment où les utilisateurs et utilisatrices sont informés que leur contenu a été modéré ou que leur compte a été pénalisé d'une manière ou d'une autre, ainsi que les motifs de ces mesures; les notifications qui sont souvent trop générales et font uniquement référence, voire pas du tout, à une politique qui aurait été violée sans autre justification¹²; l'absence de recours juridiques effectifs dans les conditions générales ainsi que la présence de clauses problématiques concernant la résolution des litiges et le choix de la loi applicable qui peuvent dissuader la plupart des utilisateurs et des utilisatrices d'intenter une action en justice en saisissant par exemple les tribunaux de leur pays.
- contournement de l'État de droit: parfois, les autorités contactent les entreprises de réseaux sociaux de manière informelle pour demander la suppression de contenus en se fondant sur les conditions générales des entreprises. Bien que les entreprises ne soient pas légalement obligées de se conformer à ces demandes, elles peuvent décider de retirer le contenu sans donner aux utilisateurs la possibilité de contester la légalité de la restriction en question devant un tribunal.

43. À cet égard, la presse et les médias d'information nécessitent des mesures réglementaires spécifiques. Ces services jouent un rôle central dans l'exercice du droit de recevoir et de communiquer des informations en ligne, et ils fonctionnent conformément à la législation et aux règles professionnelles en vigueur. Dans le cas des médias de radiodiffusion, ils sont également soumis à une surveillance réglementaire nationale. Il s'ensuit donc que lorsqu'un fournisseur de médias sociaux modère le contenu de la presse ou des médias d'information, il convient d'accorder une attention particulière aux principes de liberté et de pluralisme des médias¹³. En prenant l'EMFA comme modèle, les fournisseurs de médias sociaux devraient fournir une explication de la raison d'être de la mesure de modération en question avant sa mise en œuvre. En outre, ils devraient donner à la presse ou au média d'information en question la possibilité d'apporter un droit de réponse.

44. La modération des vidéos tournées dans les zones de conflit est particulièrement préoccupante. Compte tenu de leur caractère violent et perturbant, les entreprises de réseaux sociaux retirent généralement ces vidéos pour protéger leurs utilisateurs et utilisatrices. Au-delà de leur valeur informative, ces vidéos pourraient également servir de preuves de crimes de guerre devant un tribunal. Cette question, qui avait déjà été soulevée lors de la guerre en Syrie en 2011, est revenue sur le devant de la scène en particulier lors de la guerre d'agression de la Fédération de Russie contre l'Ukraine¹⁴. Les outils de modération de contenu automatisés qui sont utilisés semblent avoir aggravé le problème, car ils n'ont pas la capacité d'interpréter la valeur de ces séquences de guerre¹⁵. Certes, il existe des organisations privées telles que [The Syrian Archive](#) et [Mnemonic](#) qui collectent, archivent et recherchent ce type de séquences, mais les entreprises de

11. Voir: [«Content moderation and freedom of expression handbook»](#).

12. Pour plus d'informations sur la pratique du «shadow banning», voir ci-dessous au chapitre 6.

13. Pour des cas où les médias sociaux n'ont pas tenu compte de la liberté et du pluralisme des médias dans leurs activités de modération des contenus, voir [«Is big tech tampering with media content?»](#).

14. Voir: [«The difficulty of conserving social media evidence of war crimes»](#).

15. Voir: [«AI: War crimes evidence erased by social media platforms»](#).

réseaux sociaux devraient faire davantage pour éviter de les supprimer, sauf si cela est absolument nécessaire, et s'efforcer de les préserver (y compris leurs métadonnées) lorsque le contenu en question peut servir de preuve de crimes de guerre. Le problème est cependant plus large, et les entreprises de réseaux sociaux devraient chercher à éviter la suppression permanente d'éléments de preuve qui pourraient également être utilisés pour poursuivre et sanctionner d'autres crimes.

5. Options de règlement des conflits

45. Avant l'introduction du DSA, les utilisateurs et utilisatrices de l'Union européenne n'avaient que deux possibilités pour contester les mesures de modération de contenu: soit déposer une réclamation auprès du système interne de traitement des réclamations du service (si cette option était disponible), sachant que ledit service pouvait toujours prendre une décision partielle, soit emprunter la voie judiciaire en intentant une action en justice coûteuse¹⁶.

46. Le DSA a établi un cadre réglementaire complet qui comprend trois voies de règlement des conflits: le dépôt d'une réclamation auprès d'un système interne de traitement des réclamations, la saisine d'un organe de règlement extrajudiciaire des litiges et la saisine de la justice.

5.1. Système interne de traitement des réclamations

47. L'article 20 du DSA indique que les plateformes en ligne doivent mettre à la disposition des utilisateurs et utilisatrices un système qui leur permet de déposer une réclamation, par voie électronique et gratuitement, contre les décisions de modération de contenu qui leur causent un préjudice. Les personnes qui s'estiment lésées disposent d'un délai de six mois pour introduire leur réclamation à compter du jour où elles ont été informées de la décision de la plateforme. Le système interne de traitement des réclamations doit être aisément accessible, facile à utiliser et permettre ainsi que faciliter la présentation de réclamations suffisamment précises et étayées. Les plateformes en ligne, quant à elles, doivent traiter les réclamations de manière rapide, non discriminatoire, diligente et non arbitraire, et annuler dans les meilleurs délais les décisions sans fondement. Les utilisateurs et utilisatrices doivent être informés rapidement de la décision motivée et de la possibilité d'un règlement extrajudiciaire du litige ainsi que des autres voies de recours disponibles. Les décisions doivent être prises sous le contrôle d'un personnel dûment qualifié et ne pas se fonder uniquement sur des moyens automatisés.

5.2. Règlement extrajudiciaire des litiges

48. La véritable nouveauté, introduite par l'article 21 du DSA, est que les utilisateurs et utilisatrices peuvent choisir un organe certifié de règlement extrajudiciaire des litiges pour résoudre les différends qui sont liés aux décisions relatives à la modération de contenu, sachant que ces décisions ne sont pas contraignantes.

49. Les informations qui concernent la possibilité d'accéder à un organe de règlement des litiges doivent être facilement accessibles sur l'interface en ligne de la plateforme, claires et faciles à utiliser.

50. Le recours à ce type d'organe est sans préjudice du droit de l'utilisateur et de l'utilisatrice d'engager, à tout moment, une procédure judiciaire devant un tribunal. En tout état de cause, les plateformes en ligne peuvent refuser de se mettre en relation avec un organe de cette nature si un litige concernant les mêmes informations et les mêmes motifs d'illégalité ou d'incompatibilité alléguée du contenu a déjà été résolu.

51. Ces organes sont certifiés par le coordinateur des services numériques¹⁷ de l'État membre où ils sont établis. Ils doivent être impartiaux et indépendants, disposer de l'expertise nécessaire, être rémunérés d'une manière qui ne soit pas liée à l'issue de la procédure, être facilement accessibles en ligne, être capables de régler les litiges d'une manière rapide, efficace et économique, et leurs règles de procédure doivent être claires, équitables, facilement et publiquement accessibles, et conformes à la législation applicable. Les organes de règlement des litiges doivent rendre compte chaque année de leur fonctionnement au coordinateur des services numériques. Les coordinateurs des services numériques établissent tous les deux ans un rapport sur le fonctionnement des organes de règlement extrajudiciaire des litiges qu'ils ont certifiés.

16. Voir: «[Resolving content disputes outside the courtroom using the Digital Services Act](#)».

17. Les coordinateurs des services numériques sont des autorités nationales chargées de toutes les questions relatives à la supervision et à l'application du DSA dans un État membre de l'Union européenne (voir l'article 49 du DSA à ce sujet).

52. Les organes certifiés de règlement des litiges doivent mettre leurs décisions à la disposition des parties dans un délai raisonnable et au plus tard 90 jours civils après la réception de la réclamation. Dans le cas de litiges très complexes, cette période peut être portée à 180 jours.

53. En ce qui concerne les frais de procédure, la plateforme doit les couvrir quelle que soit l'issue de la décision, pour autant que l'utilisateur ou utilisatrice ait agi de bonne foi. Ce système asymétrique pourrait encourager les plateformes en ligne à réduire les coûts de leurs procédures de règlement des litiges en commettant moins d'erreurs en matière de modération de contenu, ce qui pourrait être bénéfique pour toutes les parties¹⁸.

54. Meta n'a pas attendu l'adoption du DSA pour mettre en place un organisme indépendant de règlement des conflits appelé [Oversight Board](#) ou Conseil de surveillance (CS). Ce conseil est composé de membres indépendants qui ne sont pas des employés de Meta et ne peuvent pas être révoqué·e·s par l'entreprise. Meta ne joue aucun rôle dans la sélection des nouveaux membres du CS.

55. Lorsque les utilisateurs et utilisatrices de Facebook, Instagram ou Threads ont épuisé la procédure d'appel de Meta, ces personnes peuvent saisir le CS pour contester la décision prise par Meta concernant un contenu. Meta peut également renvoyer des affaires devant cette instance. Le CS peut choisir d'annuler ou de confirmer les décisions de Meta, qui s'engage à mettre en œuvre les décisions du Conseil.

56. En outre, le CS formule des recommandations non contraignantes sur la manière dont Meta peut améliorer ses politiques de contenu, ses systèmes d'application et sa transparence globale envers les utilisateurs et utilisatrices de Facebook, d'Instagram et de Threads. Ces recommandations sont publiques et le Comité de mise en œuvre du CS, constitué de membres de celui-ci, s'appuie sur des données internes, accessibles au public, pour évaluer l'[incidence des recommandations](#).

57. Le CS publie des [rapports de transparence](#) dans lesquels figurent des informations détaillées sur l'incidence des recommandations sur les utilisateurs et utilisatrices, les décisions prises et les cas soumis au Conseil par les utilisateurs et utilisatrices. Le CS publie également des rapports annuels qui évaluent la manière dont META met en œuvre les décisions et recommandations qu'il a adoptées. L'[outil de suivi des recommandations](#) est le moyen le plus indispensable pour collecter les mesures incluses dans chaque rapport.

58. Un autre exemple, plus récent, montre comment la collaboration entre les entreprises de réseaux sociaux peut faciliter le règlement des litiges. L'[Appeals Centre Europe](#) est un nouvel organisme certifié de règlement des litiges¹⁹ qui a été lancé à la fin de l'année 2024. Il est censé régler les litiges relatifs à Facebook, TikTok et YouTube en vérifiant que la décision de l'entreprise est cohérente avec ses propres politiques de contenu, y compris les règles ou exceptions qui se réfèrent aux droits humains. L'[Appeals Centre Europe](#) comprendra une équipe interne d'experts qui sera chargée de résoudre les litiges en appliquant un contrôle humain à chaque cas. Les cas complexes seront examinés par des spécialistes ayant une expertise dans des régions, des langues ou des domaines politiques spécifiques.

59. L'[Appeals Centre Europe](#) a été créé grâce à une subvention unique du Trust du Conseil de surveillance de Meta et sera financé par les cotisations. Les entreprises de réseaux sociaux qui participeront au Appeals Centre paieront une redevance pour chaque affaire, tandis que les utilisateurs et utilisatrices qui feront appel ne paieront qu'une redevance symbolique, qui leur sera remboursée si la décision de l'[Appeals Centre](#) leur est favorable²⁰.

5.3. Recours judiciaire

60. Les droits fondamentaux à un recours juridictionnel effectif et à un procès équitable, tels que prévus à l'article 47 de la Charte des droits fondamentaux de l'Union européenne et aux articles 6 et 13 de la Convention européenne des droits de l'homme, s'appliquent aux affaires qui soulèvent des litiges liés aux activités de modération de contenu des entreprises de réseaux sociaux. Le DSA n'empêche pas les autorités judiciaires ou administratives nationales compétentes d'émettre une injonction de rétablir des contenus lorsqu'ils sont conformes aux conditions générales mais ont été supprimés par erreur (considérant 39 du DSA).

18. Voir «Settling DSA-related Disputes Outside the Courtroom: The Opportunities and Challenges Presented by Article 21 of the Digital Services Act», *Principles of the Digital Services Act*.

19. Légalement certifié au titre de l'article 21 du DSA par l'autorité irlandaise de régulation des médias Coimisiún na Meán.

20. Pour plus d'informations, voir [«EU creates 'Appeals Centre' to referee disputes with social media giants»](#).

6. Trouver le bon équilibre des droits

61. La note d'orientation du Conseil de l'Europe indique que la modération de contenu entraîne une restriction des libertés fondamentales, et qu'il incombe aux États de veiller à ce que les restrictions imposées par les régimes de régulation, d'autorégulation et de corégulation soient transparentes, justifiées, essentielles et proportionnées. Pour ce qui est des conditions générales d'utilisation, la difficulté consiste à faire la distinction entre la corégulation légitime, la pression illégitime exercée par l'État sur les sociétés privées et l'application légitime par les réseaux sociaux de leurs règles internes.

62. Les [Principes directeurs des Nations Unies relatifs aux entreprises et aux droits de l'homme](#) reconnaissent que les entreprises ont la responsabilité de respecter les droits humains, c'est-à-dire d'agir en faisant preuve d'une diligence raisonnable pour éviter de porter atteinte aux droits d'autrui et de remédier aux préjudices qui surviennent. La responsabilité d'une entreprise en matière de respect des droits humains s'applique à l'ensemble de ses activités et relations commerciales.

63. De plus, comme l'explique le considérant 45 du DSA, si la liberté contractuelle des fournisseurs de services intermédiaires devrait en principe être respectée, elle n'est toutefois pas absolue et doit respecter certaines règles concernant le contenu, l'application et le respect des conditions générales de ces fournisseurs, dans un souci de transparence, de protection des destinataires du service et de prévention de conséquences inévitables ou arbitraires.

64. En Europe, le nombre d'affaires judiciaires portant sur des conflits relatifs à la modération de contenu est en augmentation, ce qui permet de mieux appréhender la nature des questions juridiques en jeu. En général, les tribunaux ont choisi de ne pas appliquer directement le droit à la liberté d'expression pour fonder leurs décisions. Ils ont plutôt eu recours à l'application de principes juridiques fondamentaux, tels que la bonne foi, l'équité, la rupture de contrat, la protection des consommateurs, voire la protection des données.

65. Le 29 juillet 2021, la Cour fédérale de justice allemande a rendu un arrêt important dans lequel elle examinait l'équilibre des droits des réseaux sociaux et ceux des utilisateurs et utilisatrices²¹. La Cour a expliqué que Facebook, en tant qu'entreprise privée, n'était pas directement liée par des droits fondamentaux, notamment le droit à la liberté d'expression (article 5 (1)(1) de la [Loi fondamentale allemande](#)), de la même manière que l'État. En vertu du paragraphe 3 de l'article 1 de la Loi fondamentale, seuls les pouvoirs législatif, exécutif et judiciaire sont liés par les dispositions relatives aux droits fondamentaux, qui sont directement applicables. En outre, la position dominante de Facebook dans le domaine des réseaux sociaux n'entraîne pas une obligation de type étatique. En particulier, Facebook n'assume pas la fourniture de services de communication publique, ce que la Cour constitutionnelle fédérale considère comme une condition préalable pour qu'une entité privée soit liée par des droits fondamentaux au même titre que l'État. Enfin, l'entreprise Facebook est elle-même porteuse de droits fondamentaux: dans les litiges civils entre particuliers, il est essentiel d'interpréter et d'équilibrer leurs droits fondamentaux de manière à les rendre aussi effectifs que possible pour toutes les parties concernées.

66. Une interprétation similaire a été faite par un tribunal néerlandais dans un jugement du 6 octobre 2021²². Dans cette affaire, qui concernait des publications contenant des informations erronées sur la covid, le tribunal a expliqué que la Cour européenne des droits de l'homme n'accorde pas, en principe, d'effet horizontal direct aux dispositions de l'article 10 de la Convention. Même si tel avait été le cas, le tribunal aurait dû considérer qu'en l'espèce, ce droit fondamental était en concurrence avec les droits fondamentaux de LinkedIn. La Cour a statué que la politique de LinkedIn concernant la covid-19 était vague et que la société devait restaurer le compte de l'utilisateur car elle n'avait pas fourni de garanties concernant la résiliation des comptes des utilisateurs et utilisatrices. Il n'était pas nécessaire, cependant, de télécharger à nouveau les messages de désinformation en question, car LinkedIn avait des raisons valables de conclure qu'ils contenaient des informations erronées préjudiciables.

67. Il existe d'autres cas où les tribunaux se sont prononcés contre la résiliation ou la suspension du compte d'un utilisateur ou d'une utilisatrice en raison de l'imprécision ou du caractère abusif des clauses contractuelles du service. Par exemple, le 5 juin 2024, un tribunal français a jugé que la fermeture du compte d'une historienne par Facebook avait été faite en violation du contrat et que la règle appliquée par le réseau social était abusive²³. En l'espèce, l'historienne avait publié sur sa page Facebook un article concernant les activités de Daesh en Algérie. À la suite de cette publication, Facebook avait désactivé le compte de l'historienne en ne fournissant comme motif qu'un message électronique générique dans lequel la plateforme

21. Voir [BGH, III ZR 179/20](#), 29 juillet 2021 et [BGH III ZR 192/20](#). Pour une description approfondie (en anglais) de l'affaire, voir «[The Case on Facebook's Terms of Service](#)».

22. Voir [Rechtbank Noord-Holland, C/15/319230 / KG ZA 21-432, 06-10-2021](#). Voir aussi [Van Haga v. LinkedIn](#).

rappelait sa politique consistant à interdire toute menace crédible de nuire à autrui, le soutien à des organisations violentes ou les publications excessivement scandaleuses. Le tribunal a toutefois conclu que l'historienne avait dénoncé sans ambiguïté le groupe terroriste et que la simple reproduction d'un communiqué de presse de Daesh ne pouvait pas être considérée comme une approbation de leurs actions, compte tenu de la mise en contexte effectuée dans la publication. Dès lors, cette publication n'entrait pas dans le champ des actions non autorisées sur le réseau social et ne pouvait être en aucun cas considérée comme remplissant les conditions prévues par [les conditions de service de Facebook](#) pour suspendre ou résilier un compte. Par ailleurs, le tribunal a jugé que ladite clause était abusive au sens de l'article R.212-2 du [Code de la consommation](#) qui dispose que «dans les contrats conclus entre des professionnels et des consommateurs, sont présumées abusives (...) sauf au professionnel à rapporter la preuve contraire, les clauses ayant pour objet ou pour effet de (...) reconnaître au professionnel la faculté de résilier le contrat sans préavis d'une durée raisonnable»²⁴.

68. Il est important de rappeler que la modération de contenu est un processus exigeant, principalement en raison de la grande quantité de contenus qui doivent être modérés et de l'évaluation juridique complexe qui doit être effectuée au cas par cas dans de nombreuses situations. En outre, la pression sociale croissante pour lutter contre les contenus préjudiciables et les intérêts commerciaux peuvent conduire à des réactions excessives de la part des réseaux sociaux²⁵. Toute personne qui publie un contenu considéré comme choquant, controversé ou indésirable (même s'il est légal) peut être touchée par ces réactions excessives.

69. À cet égard, le «bannissement furtif» (*shadow banning*) est un type de mesure de modération de contenu qui pose un problème particulièrement épineux. Il s'agit d'une pratique par laquelle les médias sociaux déréférencent ou déclassent des contenus sans en avertir l'utilisateur ou l'utilisatrice concerné·e. Le contenu reste accessible, mais l'utilisateur ou l'utilisatrice ne sait pas qu'il est devenu invisible et que personne ne peut y accéder.

70. Il semble que le «bannissement furtif» soit utilisé par les entreprises de réseaux sociaux principalement dans le contexte de sujets controversés. Par exemple, depuis l'attaque du Hamas contre Israël le 7 octobre 2023 et la guerre qui a éclaté ensuite à Gaza, de nombreux militants propalestiniens se sont plaints du bannissement furtif du contenu qu'ils publient sur les médias sociaux. Une situation similaire s'est produite avec le contenu lié à Black Lives Matter en 2020.²⁶

71. Cette pratique, qui n'est guère compatible avec la liberté d'expression et d'information, porte gravement atteinte au droit des utilisateurs et utilisatrices de se défendre et à leur possibilité de modifier le contenu en question afin qu'il soit conforme aux conditions générales du service. Elle soulève également des questions d'équité et de protection des données. Cependant, le problème du «shadow banning» est qu'il est difficile de prouver qu'il s'est réellement produit²⁷.

72. Des cas de «shadow banning» ont déjà été qualifiés d'illégaux par les tribunaux européens. Par exemple, le 3 juin 2024, un tribunal belge a condamné Meta pour avoir interdit la page Facebook d'un homme politique²⁸. Meta reproche à cette personne d'avoir publié à plusieurs reprises des contenus qui enfreignaient les règles de Facebook relatives aux discours de haine, au soutien à des individus dangereux et à des organisations propageant la haine, ainsi qu'à l'intimidation et au harcèlement. En ce qui concerne le droit à la liberté d'expression, la Cour a statué que les décisions de Meta ne devraient pas, en principe, être évaluées au regard des critères de légalité, de légitimité et de proportionnalité définis dans l'article 10 de la Convention. Le requérant ne pouvait donc pas invoquer une violation du droit à la liberté d'expression. Cependant, Meta n'a pas agi conformément au principe d'exécution de bonne foi des accords lorsqu'elle a imposé le

23. Tribunal judiciaire de Paris – 17ème Ch. Presse-civile 5 juin 2024 / n° 21/00726. Voir aussi «Meta Meta a commis une faute contractuelle en désactivant sans préavis le compte Facebook d'une historienne ayant dénoncé les exactions de Daesh».

24. Pour un cas similaire concernant la désinformation sur la covid-19, voir [Van Haga c. YouTube](#).

25. Voir par exemple «[Remedying Overremoval](#)».

26. Voir: «[These TikTok Creators Say They're Still Being Suppressed for Posting Black Lives Matter Content](#)».

27. Voir: «[Social media users say their Palestine content is being shadow banned – here's how to know if it's happening to you](#)», «[Meta's Broken Promises – Systemic Censorship of Palestine Content on Instagram](#)» et «[Facebook, et People are accusing Instagram of shadowbanning content about Palestine](#)».

28. Les deux arrêts anonymes de la Cour d'appel dans cette affaire sont disponibles à l'adresse suivante: www.rechtbanken-tribunaux.be/sites/default/files/media/hbca/gent/files/uitreksel-inhoud-tussenarrest-24.10.2022-hof-van-beroep-gent-k-7.pdf et www.rechtbanken-tribunaux.be/sites/default/files/media/hbca/gent/files/uitreksel-inhoud-arrest-03.06.2024-hof-van-beroep-gent-k-7.pdf. Voir aussi: «[Belgian far-right MEP wins case against Meta over shadowban](#)».

bannissement furtif et n'a pas fourni de garanties procédurales suffisantes aux utilisateurs et utilisatrices lors de l'application des mesures de modération du contenu. Meta n'a pas non plus fourni d'éléments suffisants pour cette décision.

73. Un mois plus tard, le 5 juillet 2024, un tribunal néerlandais s'est prononcé contre X (ex-Twitter), qui avait fermé un compte sans prévenir l'utilisateur²⁹. Le requérant avait publié un contenu critiquant la Commission européenne d'avoir diffusé des informations trompeuses concernant sa proposition de règlement visant à prévenir et combattre les abus sexuels commis contre des enfants. Selon X, les restrictions avaient peut-être été déclenchées par ses systèmes automatisés. En janvier 2024, le requérant a été informé que sa publication avait été associée à tort à l'exploitation sexuelle des enfants et que la restriction avait été levée. La partie intéressante de cette affaire concerne l'argumentation de X selon laquelle ses conditions générales lui réservaient le droit de limiter l'accès à divers aspects et fonctionnalités de son service, et que ses obligations à l'égard de l'utilisateur avaient été remplies puisque le requérant avait eu accès à d'autres fonctionnalités essentielles. Le tribunal a estimé que la clause permettant à Twitter de suspendre ou de résilier l'accès à son service payant à tout moment et sans raison était contraire à la [Directive de l'Union européenne concernant les clauses abusives dans les contrats conclus avec les consommateurs](#). Il a également jugé que X avait agi en violation des articles 12 et 17 de la loi sur la protection des données étant donné que ses deux premières réponses à la demande d'information du requérant avaient été trop vagues et n'explicitaient pas les raisons exactes de la restriction, et que le centre d'assistance de X ne permettait pas aux utilisateurs d'établir un dialogue effectif avec l'entreprise.

7. Améliorer la modération de contenu

74. Pour éviter les pratiques de modération de contenu susceptibles de violer les droits des utilisateurs, je proposerai quatre lignes directrices sur lesquelles les autorités publiques et les entreprises de réseaux sociaux pourraient travailler ensemble: intégrer directement les principes des droits fondamentaux dans les conditions générales; améliorer les conditions générales pour renforcer leur clarté; fournir aux modérateurs et modératrices humains une formation appropriée et de bonnes conditions de travail; et utiliser l'intelligence artificielle (IA) d'une manière plus efficace. Ce chapitre se termine par une analyse du rôle des algorithmes de recommandation dans la promotion de la liberté d'expression et de la pluralité et de la diversité des points de vue.

7.1. Intégrer les principes des droits fondamentaux dans les conditions générales

75. Selon le Rapporteur spécial des Nations Unies sur la promotion et la protection du droit à la liberté d'opinion et d'expression³⁰, «les conditions d'utilisation ne devraient pas reposer sur une approche discrétionnaire fondée sur les besoins génériques et les intérêts égoïstes de la 'communauté'». Ils devraient être repensés afin que les utilisateurs «exposent leurs opinions, s'expriment librement et accèdent à toutes sortes d'informations d'une manière compatible avec le droit des droits de l'homme». Les entreprises devraient faire figurer dans leurs conditions générales les principes du droit des droits humains afin de garantir que la modération de contenu est conforme «aux principes de légalité, de nécessité et de légitimité par lesquels les États sont liés lorsqu'ils réglementent la liberté d'expression».

76. Comme mentionné ci-dessus, l'article 14 du DSA prévoit que les fournisseurs de services intermédiaires doivent agir en tenant dûment compte des droits et des intérêts légitimes de toutes les parties impliquées, et notamment des droits fondamentaux des utilisateurs et utilisatrices, tels qu'ils sont consacrés par la Charte des droits fondamentaux de l'Union européenne. L'approche fondée sur les droits humains est également celle que les organes du Conseil de l'Europe encouragent.

77. L'inscription des droits fondamentaux dans les conditions générales renforcera leur applicabilité. Toutefois, le problème de l'équilibre entre la liberté des réseaux sociaux et la liberté d'expression des utilisateurs et utilisatrices reste entier³¹. À cet égard, les entreprises de réseaux sociaux devraient s'efforcer de clarifier la manière dont leurs politiques de modération de contenu sont destinées à concilier ces deux libertés lorsqu'elles peuvent être en conflit.

29. Voir les deux jugements du tribunal de district d'Amsterdam à <https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RBAMS:2024:3980> et <https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RBAMS:2024:4019>. Voir également «[NL] District Court of Amsterdam rules that X has violated the DSA and the GDPR by 'shadowbanning' its user».

30. Rapport du Rapporteur spécial sur la promotion et la protection du droit à la liberté d'opinion et d'expression, 6 avril 2018.

31. Voir par exemple: «Using Terms and Conditions to apply Fundamental Rights to Content Moderation».

7.2. Clarté des conditions générales

78. Les conditions générales doivent expliquer la nature des contenus autorisés sur le service et quelles sont les conséquences de la publication d'un contenu illégal ou contraire aux règles de l'entreprise de réseaux sociaux. Cette explication doit être complète, précise et présentée de manière que tout utilisateur et toute utilisatrice (et pas seulement les utilisateurs et utilisatrices avertis) puisse la comprendre.

79. Le 9 août 2023, l'Ofcom, l'autorité britannique de régulation des communications, a publié un [rapport sur les politiques relatives aux utilisateurs et utilisatrices des plateformes de partage de vidéos](#). Le rapport met en lumière les approches adoptées par six plateformes de partage de vidéos (BitChute, Brand New Tube, OnlyFans, Snap, TikTok et Twitch) pour élaborer et mettre en œuvre des conditions générales qui protègent les utilisateurs et utilisatrices, tout en citant des exemples de bonnes pratiques. Les principales recommandations sont les suivantes:

- Les conditions générales doivent être claires et faciles à comprendre:
 - les conditions générales doivent être faciles à trouver et à consulter; pour les services qui comptent un grand nombre d'enfants parmi leurs utilisateurs et utilisatrices, cela pourrait signifier une section distincte expliquant comment les enfants sont protégés sur la plateforme;
 - elles doivent être rédigées de sorte qu'elles soient comprises par le plus grand nombre d'utilisateurs et d'utilisatrices possible, y compris les enfants et les personnes qui n'ont pas de compétences avancées en lecture;
 - il faut examiner les techniques qui permettent de mesurer et d'améliorer la participation des utilisateurs et des utilisatrices et leur compréhension des conditions générales.
- Les conditions générales doivent protéger tous les utilisateurs et toutes les utilisatrices contre les contenus préjudiciables. Voici quelques exemples de bonnes pratiques:
 - couvrir les différents types de contenus soumis à des restrictions qui sont susceptibles de porter préjudice aux enfants;
 - clarifier les contenus autorisés et interdits d'une manière compréhensible pour un enfant;
 - en cas d'utilisation de classifications relatives à l'âge ou à la sensibilité, expliquer clairement aux utilisateurs et aux utilisatrices quels types de contenus devraient être classés parmi les contenus inappropriés pour les enfants.
- Les conditions générales devraient indiquer clairement les contenus qui ne sont pas autorisés et quelles sont les conséquences en cas de non-respect des règles. Pour ce faire, il faut notamment:
 - définir les contenus autorisés et interdits sur la plateforme (sauf si des raisons exceptionnelles justifient de ne pas le faire);
 - expliquer toutes les mesures qui pourraient être prises pour le fournisseur de service si un utilisateur ou une utilisatrice enfreint les règles de la plateforme de partage de vidéos.

7.3. Formation adéquate et bonnes conditions de travail pour les modérateurs et modératrices

80. Toutes les entreprises de réseaux sociaux utilisent (à des degrés divers)³² des modérateurs et des modératrices humains qui prennent des décisions sur le contenu généré par les utilisateurs et utilisatrices.

81. Les modérateurs et les modératrices humains ont besoin d'une formation appropriée pour acquérir une connaissance suffisante de la loi et des lignes directrices de l'entreprise qu'ils doivent appliquer. Le rapport de l'Ofcom mentionné ci-dessus recommande d'établir des définitions complètes de la terminologie clé, illustrées par des études de cas, ainsi que des exemples de contexte, d'exceptions et de considérations concernant le matériel préjudiciable. Il suggère également de fournir des études de cas audio et/ou visuelles de contenus préjudiciables et de fournir des conseils détaillés sur la manière dont les modérateurs doivent répondre à l'évolution des préjudices et des comportements en ligne qui émergent dans un contexte de crise. En outre, les conditions générales et les directives adressées aux modérateurs et modératrices devraient faire l'objet d'une révision permanente et leur efficacité devrait être testée.

32. Voir: «Content moderation: Key facts to learn from Facebook, Instagram, X and TikTok transparency reports» (Modération en ligne: ce que nous apprennent les rapports de transparence de Facebook, Instagram, TikTok).

82. Les modérateurs et les modératrices humains ont également besoin de bonnes conditions de travail, et surtout d'un soutien psychologique. Les problèmes liés aux conditions de travail comprennent les bas salaires, la sous-traitance, les accords de non-divulgaration qui peuvent empêcher les modérateurs et les modératrices de parler du contenu qu'ils et elles modèrent et de ses effets sur leur santé mentale, ainsi que les clauses de non-responsabilité relatives au syndrome de stress post-traumatique qui doivent être signées par les modérateurs et modératrices, reconnaissant les risques du métier pour la santé mentale³³.

83. En Allemagne, [en 2023, des modérateurs et modératrices de contenu de réseaux sociaux ont publié un manifeste](#) appelant à la fin de l'exploitation dans la modération de contenu. Les signataires ont exigé des changements immédiats à l'échelle de leur secteur, notamment: de meilleurs salaires et avantages; une prime de risque équivalant à au moins 35 % du salaire annuel de modérateur; l'accès à des soins de santé mentale appropriés 24 heures sur 24; la fin de la culture du secret et de l'intimidation, y compris la dissolution des accords de non-divulgaration précédemment signés; le soutien des entreprises de réseaux sociaux aux modérateurs et modératrices de contenu pour qu'ils et elles puissent s'organiser collectivement, négocier et adhérer à un syndicat; la fin de l'externalisation du travail essentiel pour la sécurité qu'est la modération de contenu; la suppression de toute surveillance oppressive et déraisonnable et de la gestion algorithmique; l'égalité des rémunérations indépendamment des expériences passées ou du pays de résidence.

84. En ce qui concerne la santé mentale des modérateurs et des modératrices, une [étude réalisée en 2023 sur les effets psychologiques de la modération de contenu sur les modérateurs et modératrices](#) a conclu que celles et ceux exposés à du matériel d'abus sexuel d'enfants présentaient une série de symptômes qui rentrent dans la catégorie des symptômes associés au stress post-traumatique³⁴ et au stress traumatique secondaire³⁵, comparables à ceux dont peuvent souffrir les professionnel·le·s travaillant dans les services d'urgence ou dans les activités de soins, comme les travailleuses et les travailleurs sociaux. L'étude suggère que les entreprises qui emploient des modérateurs et modératrices devraient s'inspirer de ces professions et offrir à leurs employé·e·s des programmes de psycho-éducation et une prise en charge qui tienne compte des traumatismes.

85. À cet égard, il est important que les problèmes de santé mentale causés par le travail des modérateurs et modératrices de contenu soient reconnus comme des maladies liées au travail et que les employeurs et le système judiciaire en soient conscients. Ces dernières années, cette question a pris de l'ampleur, les modérateurs et modératrices prenant la parole en dépit des accords de non-divulgaration et de la pression interne. Par exemple, en 2020, Facebook a accepté de payer 52 millions USD pour régler un recours collectif intenté par des modérateurs et modératrices au sujet des problèmes de santé mentale qu'ils et elles avaient développés dans le cadre de leur travail. Plus récemment, en janvier 2024, un tribunal de Barcelone a jugé que les problèmes psychologiques d'un modérateur de contenu travaillant pour un sous-traitant de Meta constituaient une maladie professionnelle. Dans cette affaire, le tribunal a conclu que le facteur de stress professionnel était le seul, exclusif et incontestable déclencheur des dommages psychologiques causés à l'employé, qui était exposé à des contenus pénibles tels que des automutilations, des décapitations de personnes civiles tuées par des groupes terroristes, des tortures infligées à des personnes ou des suicides³⁶.

7.4. Utiliser l'intelligence artificielle de manière efficace

86. L'un des principaux problèmes liés à la modération de contenu sur internet est la quantité énorme de contenus qui y circulent. Selon les estimations du Forum économique mondial, d'ici 2025, environ 463 exaoctets seront créés chaque jour³⁷. À cet égard, l'IA peut contribuer de manière significative à automatiser la modération de contenu sur les services en ligne. Les systèmes d'IA peuvent analyser et classer rapidement des masses de contenus potentiellement préjudiciables (y compris les contenus publiés en temps réel), ce que les êtres humains ne peuvent pas faire. Ils peuvent aussi filtrer les contenus les plus préjudiciables afin que les modérateurs et modératrices humains n'y soient pas exposés. Ils peuvent fonctionner de manière autonome ou être combinés à une intervention humaine dans les cas où la machine seule n'est pas en mesure de faire le travail.

33. Voir: [«Facebook moderator: 'Every day was a nightmare'»](#) et [«Facebook and YouTube moderators sign PTSD disclosure»](#).

34. Voir: [«Post-traumatic stress disorder \(PTSD\)»](#).

35. Voir: [«Secondary Trauma explained»](#).

36. Voir: [«Primera sentencia contra la subcontrata de Meta en Barcelona por daños mentales a un empleado»](#).

37. Voir: [«How much data is generated each day?»](#)

87. Une [étude réalisée en 2019 par Cambridge Consultants](#) explique les trois façons dont les technologies d'IA peuvent avoir une incidence considérable sur les flux de travail liés à la modération de contenu:

- l'IA peut signaler les contenus pour qu'ils soient examinés par des êtres humains, augmentant ainsi la précision de la modération;
- l'IA peut créer des données d'entraînement pour améliorer la performance de la modération à priori;
- l'IA peut assister les modérateurs et modératrices en augmentant leur productivité et en réduisant les effets potentiellement préjudiciables de la modération de contenu sur eux et elles-mêmes.

88. Toutefois, la modération de contenu par l'IA n'est pas une solution parfaite. Par exemple, un [rapport de 2022 de la Commission fédérale du commerce des États-Unis au Congrès](#) a invité les responsables politiques à être très prudents quant à l'utilisation de l'IA comme solution politique, car les outils d'IA peuvent être inexacts, biaisés, discriminatoires de par leur conception et peuvent encourager le recours à des formes de surveillance commerciale de plus en plus invasives. Plus important encore, ces programmes ont des difficultés à comprendre le contexte (humour, sarcasme, références culturelles), ce qui peut conduire à des interprétations erronées du contenu des utilisateurs et utilisatrices, et ils doivent être formés en permanence pour s'adapter aux formes changeantes de contenu préjudiciable³⁸.

89. Une autre question importante est celle de la modération de contenus générés par l'IA. Les technologies de l'IA permettent de créer toutes sortes de contenus médiatiques (non seulement des textes, mais aussi des documents audio et vidéo) qu'il est pratiquement impossible de distinguer de la réalité. Par conséquent, les utilisateurs et les utilisatrices doivent être informés lorsqu'ils et elles font face à des textes, des images ou des sons générés par l'IA, car ce type de contenus peut être particulièrement trompeur et contenir entre autres dangers des informations délibérément fausses et des propos haineux.

90. L'article 8 de 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](#) exige que les Parties à la Convention «adoptent ou maintiennent» des mesures pour veiller à ce que les exigences de transparence et de contrôle adaptées aux contextes et aux risques spécifiques soient en place en ce qui concerne les activités au sein du cycle de vie des systèmes d'intelligence artificielle, y compris en ce qui concerne l'identification de contenu généré par des systèmes d'intelligence artificielle».

91. L'article 50, paragraphe 2, de la [loi européenne sur l'intelligence artificielle](#) prévoit que les contenus de synthèse de type audio, image, vidéo ou texte soient marqués dans un format lisible par machine et identifiables comme ayant été générés ou manipulés par une IA. Les fournisseurs doivent veiller à ce que leurs solutions techniques soient aussi efficaces, interopérables, solides et fiables que la technologie le permet, compte tenu des spécificités et des limites des différents types de contenus, des coûts de mise en œuvre et de l'état de la technique généralement reconnu. Toutefois, cette obligation ne s'applique pas aux systèmes d'IA qui remplissent une fonction d'assistance pour la mise en forme standard ou qui ne modifient pas de manière substantielle les données d'entrée fournies par le déployeur ou leur sémantique, ou lorsque leur utilisation est autorisée par la loi à des fins de prévention ou de détection des infractions pénales, d'enquêtes ou de poursuites en la matière.

92. En septembre 2024, le [Conseil de surveillance de Meta a publié les résultats d'une enquête](#) sur l'application des politiques de contenu par les algorithmes d'IA et les techniques d'automatisation. L'enquête a débouché sur une série de recommandations, qui sont présentées ci-dessous:

- les plateformes devraient centrer leurs politiques sur la détection de l'absence de consentement des personnes visées par les hypertrucages ou «deepfakes». La génération ou la manipulation par l'IA devrait être considérée comme un signal indiquant que ces images pourraient être non consenties.
- l'automatisation devrait permettre aux utilisateurs et utilisatrices de mieux comprendre les politiques et d'éviter la suppression erronée de leur propre contenu, notamment au moyen de notifications informatives aux utilisateurs et utilisatrices. En outre, les internautes devraient pouvoir fournir le contexte de leur publication, que les modérateurs et modératrices de contenu, humains ou automatiques, n'ont peut-être pas interprété correctement, comme la satire, la sensibilisation ou la condamnation.
- les avantages des nouveaux modèles d'IA générative devraient être partagés équitablement entre les bases mondiales d'utilisateurs et utilisatrices des entreprises de réseaux sociaux, et non limités aux pays anglophones ou aux marchés occidentaux.

38. Voir: [«The Impact of AI and Machine Learning on Content Moderation: Advancements and Challenges»](#).

- les systèmes automatisés de modération devraient être soumis à une évaluation rigoureuse et continue de leurs performances, en particulier pour les utilisateurs et utilisatrices les plus vulnérables et les plus exposés. Les nouveaux modèles ne devraient pas accentuer les préjugés sociétaux existants qui peuvent avoir un impact préjudiciable sur les groupes marginalisés et d'autres individus lorsqu'ils sont déployés.
- des experts mondiaux en matière de droits humains, de liberté d'expression et d'éthique devraient être consultés avant de déployer de nouveaux outils de modération de contenu par l'IA. Leurs recommandations en matière d'atténuation des risques et d'autres garde-fous devraient être intégrées à leur conception.
- des chercheurs indépendants du monde entier devraient avoir accès aux données nécessaires pour évaluer l'impact de la modération algorithmique, de la gestion des flux et des outils d'IA sur le contenu généré par les utilisateurs.
- les réseaux sociaux devraient apposer des étiquettes indiquant aux utilisateurs et utilisatrices que le contenu est considérablement modifié et pourrait les induire en erreur. Il est également indispensable d'allouer des ressources suffisantes à l'examen humain qui soutient ce processus.

93. Il existe plusieurs options qui permettent d'informer les utilisateurs et utilisatrices sur les contenus générés par l'IA, notamment l'étiquetage de ces contenus³⁹, l'utilisation d'outils automatisés de vérification des faits⁴⁰, l'analyse technico-légale numérique⁴¹ et, en particulier, les techniques de tatouage numérique (*watermarking*).

94. Le tatouage numérique est un processus qui consiste à intégrer dans la sortie d'un modèle d'IA un signal unique et reconnaissable (un filigrane ou *watermark* en anglais) qui sert à identifier le contenu comme ayant été généré par l'IA⁴². Le tatouage numérique présente un certain nombre d'avantages: il permet d'authentifier le contenu et de contrôler les données, d'identifier l'auteur et de protéger les droits d'auteur, ainsi que d'empêcher la diffusion de fausses informations générées par l'IA. Toutefois, à ce jour, il présente un certain nombre de limites et d'inconvénients, notamment:

- le manque d'interopérabilité des différents systèmes de tatouage numérique;
- les difficultés techniques liées à l'ajout d'un filigrane pour marquer les textes;
- le fait que les filigranes peuvent être manipulés, supprimés ou modifiés.

7.5. Le rôle des systèmes de recommandation

95. Comme le rappelle le considérant 70 du DSA, les plateformes en ligne utilisent des algorithmes pour suggérer, classer et hiérarchiser les informations, en les distinguant par le texte ou d'autres représentations visuelles, ou en organisant de toute autre manière les informations fournies par les destinataires. Ces systèmes de recommandation ont un impact significatif sur la recherche et l'interaction des utilisateurs et utilisatrices avec l'information en ligne et jouent un rôle important dans l'amplification de certains messages, la diffusion virale de l'information et la stimulation du comportement en ligne. Il est donc essentiel que les utilisateurs et utilisatrices soient correctement informés de la façon dont les systèmes de recommandation influencent la manière dont les informations sont affichées et peuvent influencer la manière dont les informations leur sont présentées. Comme expliqué ci-dessus, l'article 27 de la DSA a introduit un certain nombre de règles concernant la transparence des systèmes de recommandation.

96. L'Assemblée a déjà abordé la question de la transparence algorithmique. En particulier, son rapport intitulé «Médias sociaux: créateurs de liens sociaux ou menaces pour les droits humains?»⁴³ rappelle que la sélection algorithmique conduit à un manque d'exposition à diverses sources d'information, un phénomène connu sous le nom de «bulle de filtrage» ou «chambre d'écho», et contribue à la radicalisation et à la montée de l'esprit partisan dans la société. Les algorithmes peuvent toutefois être conçus et mis en œuvre pour encourager la pluralité et la diversité des points de vue, des attitudes et des opinions. Idéalement, les entreprises devraient faire appel à une évaluation et à un audit externes afin de déterminer si leurs algorithmes ne sont pas biaisés et s'ils favorisent la pluralité ou la diversité des faits, des points de vue et des opinions. Même s'il n'existe pas de mécanisme permettant de rendre cette recommandation obligatoire, un

39. Voir: «[Misinformation warning labels are widely effective: A review of warning effects and their moderating features](#)».

40. Voir: «[Emerging technologies and automated fact-checking: tools, techniques and algorithms](#)».

41. Voir: «[Machine learning in digital forensics: a systematic literature review](#)».

42. Voir: «[Generative AI and watermarking](#)».

43. Doc. 14844.

«label de bonnes pratiques» pourrait être attribué aux opérateurs internet dont les algorithmes sont conçus pour favoriser la sélection de contenus pluralistes, permettant ainsi une exposition idéologiquement transversale.

7.6. Résumé

97. La modération de contenu est une question complexe qui impose de prendre des décisions en une fraction de seconde sur des millions de contenus. En outre, la forte pression qui est exercée sur les entreprises de réseaux sociaux pour qu'elles mettent fin aux contenus illégaux et préjudiciables et qu'elles coopèrent avec les autorités publiques, par exemple dans la lutte contre la propagande de guerre, la désinformation et les discours de haine, peut les amener à être trop prudentes dans la modération de contenu et à supprimer des éléments qui sont légaux. Il est donc essentiel que toute intervention réglementaire dans ce domaine n'ait pas de conséquences imprévues pour la liberté d'expression tout en tenant dûment compte des droits et des intérêts des entreprises de réseaux sociaux.

98. Compte tenu de ces conclusions, je propose une série de mesures concrètes dans le projet de résolution.