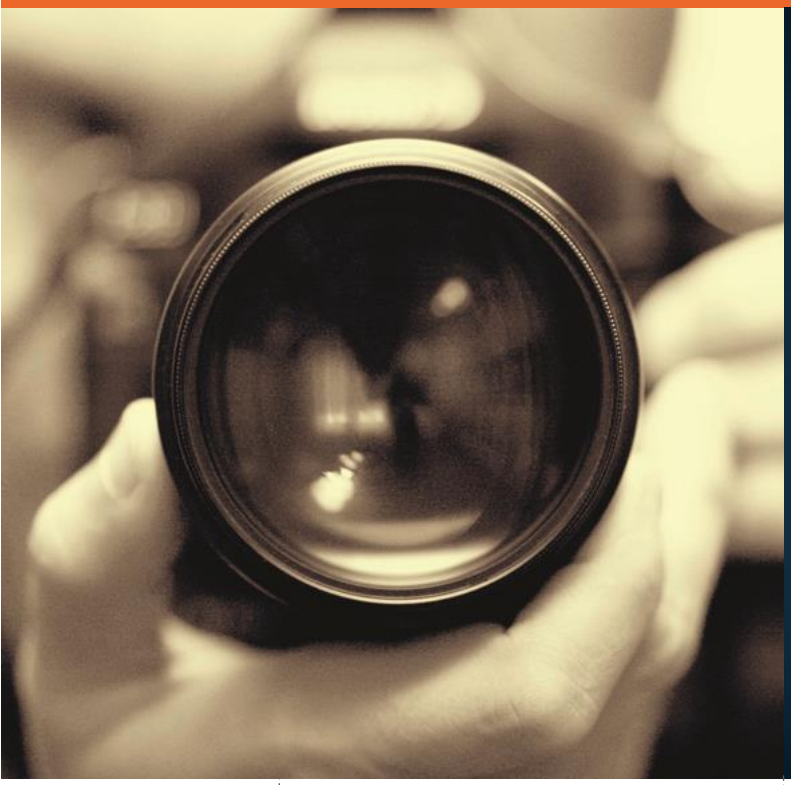
MODÉRATION DE CONTENU

Meilleures pratiques en vue de la mise en place de cadres juridiques et procéduraux efficaces pour les mécanismes d'autorégulation et de corégulation de la modération de contenu



Note d'orientation

Adoptée par le Comité directeur sur les médias et la société de l'information (CDMSI)

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

MODÉRATION DE CONTENU

Meilleures pratiques en vue de la mise en place de cadres juridiques et procéduraux efficaces pour les mécanismes d'autorégulation et de corégulation de la modération de contenu

Note d'orientation
adoptée par le Comité directeur sur les médias et la société de l'information (CDMSI) lors de sa 19me réunion plénière, 19-21 mai 2021

Conseil de l'Europe

Mises en page et page de garde :
Service de la société de l'information
Conseil de l'Europe

Images : Shutterstock

Cette publication n'a pas été revue par l'unité éditoriale du SPDP afin de corriger les erreurs typographiques et grammaticales.

Conseil de l'Europe, juin 2021

Contents

Note d'orientation

sur les meilleures pratiques en vue de la mise en place de cadres juridiques et procéduraux efficaces pour les mécanismes d'autorégulation et de corégulation de la modération de contenu

Introduction

1. La modération de contenu (envisagée ici au sens large du terme, qui comprend également l'organisation (ou « curation ») de contenu) est de plus en plus employée pour répondre à divers problèmes provenant de l'activité en ligne des utilisateurs. Elle présente des problématiques particulières car, du fait des technologies en mutation, des possibilités que celles-ci offrent et de l'évolution constante du comportement humain dans l'environnement en ligne, « il n'existe pas de résultat final fixe à atteindre en matière de modération de contenu, avec des règles ou des formes de régulation stables ; cela sera toujours fonction de contestations, d'itérations et des évolutions technologiques »[1].

2. L'autorégulation et la corégulation, souvent présentées comme deux approches bien distinctes de la modération de contenu, ne sont en fait que des degrés variables d'une même échelle, qui va d'une approche purement autorégulatoire à une extrémité à une approche de pure réglementation à l'autre extrémité, selon le degré d'implication de l'État dans le processus[2].

3. La corégulation se caractérise par un degré de mobilisation plus important de l'État et est souvent liée à la réalisation d'objectifs de politiques publiques, alors que l'autorégulation est généralement introduite de manière indépendante par des intermédiaires d'internet, pour des raisons directement liées à leur modèle commercial (voir, ci-après, « à des fins commerciales »).

4. Cependant, quel qu'en soit le degré, la modération de contenu fait inévitablement intervenir des considérations relatives aux droits humains. Les États devraient garder à l'esprit que les obligations positives et négatives qui leur incombent en matière de droits humains, y compris en ce qui concerne les droits à la liberté d'expression, au respect de la vie privée, à la liberté de réunion et d'association, à l'égalité et à la non-discrimination, ainsi que le droit à un recours effectif, résultent également de la modération de contenu réalisée dans des cadres autorégulatoires ou corégulatoires[3].

5. La présente note d'orientation a pour objet de fournir des conseils pratiques aux États membres du Conseil de l'Europe afin que l'élaboration de politiques, la régulation et le recours à la

[1] Douek Evelyn, « The limits of international law in content moderation », *UCI Journal of International, Transnational, and Comparative Law*, décembre 2020, p. 9, disponible à l'adresse https://ssrn.com/abstract=3709566, consulté le 21 janvier 2021.

[2] Marsden C.T. (2012) *Internet co-regulation and constitutionalism: Towards European judicial review International Review of Law Computers & Technology*, 26(2):211-228, disponible à l'adresse www.researchgate.net/publication/254294662_Internet_co-regulation_and_constitutionalism_Towards_European_judicial_review, consulté le 10 juin 2020.

[3] *Peck c. Royaume-Uni* (arrêt), nº 44647/98, 28 janvier 2003, par. 108 et 109.

modération de contenu en ligne soient conformes aux obligations qui leur incombent dans le domaine des droits humains au titre 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[4].

6. Cette note d'orientation vise donc à définir les meilleures pratiques à adopter en vue de disposer de cadres juridiques et procéduraux efficaces pour les mécanismes d'autorégulation et de corégulation de modération de contenu.

7. La note d'orientation fait fond sur un ensemble de travaux remarquables déjà menés par différentes institutions sur divers aspects de la modération de contenu, qui sont référencés dans l'exposé des motifs figurant en annexe. Il convient en particulier de noter les travaux du Conseil de l'Europe[5], des Nations Unies[6, 7, 8], des milieux universitaires[9, 10, 11] et de diverses ONG[12, 13, 14, 15, 16, 17]. L'exposé des motifs vise à accompagner et expliquer les dispositions, concepts et termes qui figurent dans la note d'orientation et à servir de référence fréquente lors de la mise en œuvre.

Considérations générales aux fins de l'élaboration de politiques adéquates

8. Les États doivent garder à l'esprit que la modération de contenu fait naître des questions complexes et non résolues en matière de juridiction. Il convient de tenir dûment compte de la

[4] Voir Principes directeurs relatifs aux entreprises et aux droits de l'homme - mise en œuvre du cadre de référence « protéger, respecter et réparer » des Nations Unies et Recommandation CM/Rec(2016)3 du Comité des Ministres aux États membres sur les droits de l'homme et les entreprises.

[5] Voir, par exemple, Recommandation CM/Rec(2018)2 du Comité des Ministres aux États membres sur les rôles et les responsabilités des intermédiaires d'internet et Recommandation CM/Rec(2020)1 du Comité des Ministres aux États membres sur les impacts des systèmes algorithmiques sur les droits de l'homme.

[6] Conseil des droits de l'homme de l'ONU, Rapport du rapporteur spécial sur la promotion et la protection du droit à la liberté d'opinion et d'expression, Frank La Rue, 2011, disponible à l'adresse https://documents-dds-ny.un.org/doc/UNDOC/GEN/G11/132/02/PDF/G1113202.pdf?OpenElement, consulté le 23 septembre 2020.

[7] David Kaye (2019) « A New Constitution for Content Moderation », disponible à l'adresse https://onezero.medium.com/a-new-constitution-for-content-moderation-6249af611bdf, consulté le 23 septembre 2020; Rapport du Rapporteur spécial sur la promotion et la protection du droit à la liberté d'opinion et d'expression, 6 avril 2018, A/HRC/38/35, disponible à l'adresse https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/096/73/PDF/G1809673.pdf?OpenElement, consulté le 23 avril 2021.

[8] Broadband Commission for Sustainable Development, rapport intitulé « Balancing Act: Countering Digital Disinformation while respecting Freedom of Expression », 2020, disponible à l'adresse https://en.unesco.org/publications/balanceact, consulté le 23 avril 2021.

[9] Voir www.ivir.nl/publications/technology-and-law/

[10] Luca Belli, Nicolo Zingales (2017) *Platform Regulations: How platforms are regulated and how they regulate us*, ISBN 978-85-9597-014-4, disponible à l'adresse https://direitorio.fgv.br/publicacoes/platform-regulations-how-platforms-are-regulated-and-how-they-regulate-us, consulté le 19 février 2021.

[11] Voir https://cyberlaw.stanford.edu/focus-areas/intermediary-liability

[12] AccessNow (2020) « 26 recommendations on content governance », disponible à l'adresse www.accessnow.org/cms/assets/uploads/2020/03/Recommendations-On-Content-Governance-digital.pdf, consulté le 23 septembre 2020.

[13] Article 19 (2018) « Side-stepping Rights: Regulating Speech by Contract », disponible à l'adresse www.article19.org/wp-content/uploads/2018/06/Regulating-speech-by-contract-WEB-v2.pdf, consulté le 4 janvier 2020.

[14] Voir santaclaraprinciples.org, consulté le 23 septembre 2020.

[15] Voir platformregulation.eu, consulté le 23 septembre 2020.

[16] Voir https://edri.org/?s=content+moderation, consulté le 23 septembre 2020.

[17] Meedan (2018) « Content Moderation Toolkit », disponible à l'adresse https://meedan.com/reports/content-moderation-toolkit/, consulté le 19 janvier 2020.

nature transfrontalière de l'environnement en ligne dans toute décision portant sur la régulation de la modération de contenu, ainsi que lors des évaluations d'impact[18, 19].

9. Du fait de l'évolution rapide et constante de l'environnement en ligne, les politiques de modération de contenu doivent faire l'objet d'un examen régulier. Les États doivent veiller à ce que les politiques et les cadres de régulation pertinents soient supervisés de façon adéquate et soient adaptés en temps voulu.

10. La modération de contenu est utilisée pour répondre à un vaste ensemble de problèmes relevant des politiques publiques, qui comprennent aussi bien diverses formes de comportement criminel que des contenus qui ne sont pas illégaux, ainsi que des contenus modérés par des intermédiaires d'internet à des fins commerciales (comme le traitement des courriers électroniques publicitaires non sollicités (ou « spam »)). Il est donc fondamental, pour élaborer des politiques adéquates dans ce domaine, de bien comprendre et cerner la nature des contenus visés et la responsabilité des intermédiaires d'internet dans le processus décisionnel.

11. De même, il importe de tenir compte, lors de l'élaboration des politiques, des différentes problématiques que présentent l'autorégulation et la corégulation dans différents contextes. Ces problématiques varient, par exemple, selon la nature et le volume du contenu hébergé et sa portée, et ne sont pas les mêmes pour les intermédiaires d'internet que pour les médias.

12. Les États ont à cet égard diverses obligations positives et négatives. Ils doivent mettre en place des cadres de régulation suffisamment élaborés pour que la modération de contenu garantisse aux utilisateurs d'internet, y compris aux victimes de contenus illégaux, l'exercice et la jouissance des droits humains. Les États doivent protéger les droits à la liberté d'expression, au respect de la vie privée, à la liberté de réunion et d'association, à l'égalité et à la non-discrimination, le droit à un recours effectif et les autres droits humains de toute personne relevant de leur juridiction dès lors que la modération de contenu a une incidence sur ces droits.

13. Conformément au principe de proportionnalité, qui est défini de manière approfondie dans la jurisprudence de la Cour européenne des droits de l'homme, toute limitation nécessaire des droits humains doit être la moins restrictive possible[20]. Au-delà du choix binaire entre la suppression ou le maintien d'un contenu, la modération de contenu offre une gamme d'outils (tels que la rétrogradation temporaire ou permanente, la démonétisation ou le signalement d'un contenu jugé problématique) qui devraient être pris en compte par les décideurs politiques et utilisés en tenant dûment compte de la nature du contenu visé.

---

[18] Par exemple, un contenu interdit dans une zone de juridiction donnée peut être supprimé à l'échelle mondiale, par l'application de décisions de justice dans ladite zone ou peut, parfois, n'être restreint que dans cette zone. (Google supprime dans le monde entier les contenus ayant fait l'objet de plaintes conformes à la procédure établie par la loi américaine *Digital Millennium Copyright Act*, mais limite aux recherches effectuées dans le cadre de ses activités dans l'UE les restrictions imposées en application de l'arrêt de la Cour de justice de l'UE dit du « droit à l'oubli » (affaire C-131/12)).

[19] De même, les règles en matière de responsabilité imposées dans de plus vastes zones de juridiction, ou dans celles où les intermédiaires sont établis, peuvent avoir un impact sur la liberté d'expression et le droit de recevoir et de diffuser des informations au niveau mondial. Voir, par exemple, Dan Jerker B. Svantesson, « Internet and Jurisdiction Global Situation Report 2019 », disponible à l'adresse www.internetjurisdiction.net/uploads/pdfs/GSR2019/Internet-Jurisdiction-Global-Status-Report-2019_web.pdf, consulté le 28 septembre 2020, et Internet & Jurisdiction Project (2020) « I&J Outcomes: Mappings of Key Elements of Content Moderation », disponible à l'adresse https://www.internetjurisdiction.net/news/i-j-outcomes-mappings-of-key-elements-of-content-moderation, consulté le 28 septembre 2020.

[20] Voir, par exemple, *Autronic AG c. Suisse*, n° 12726/87, 22 mai 1990, par. 61; *Weber c. Suisse*, n° 11034/84, 22 mai 1990, par. 47; *Barthold c. Allemagne*, n° 8734/79, 25 mars 1985, par. 58; *Klass et autres c. Allemagne*, n° 5029/71, 6 septembre 1978, par. 42; *Sunday Times c. Royaume-Uni*, n° 6538/74, 26 avril 1979, par. 65; *Observer and Guardian c. Royaume-Uni,* n° *13585/88*, 26 novembre 1991, par. 71.

14. En ce qui concerne la modération de contenu mise en œuvre pour répondre à des objectifs de politiques publiques, les États doivent ainsi :

    a. avoir bien défini la nature du ou des problème(s) à régler ;

    b. assurer la prévisibilité des mesures mises en œuvre ;

    c. veiller à ce que les cadres juridiques et de régulation n'entraînent pas de conformité excessive ou de mise en œuvre discriminatoire ;

    d. recenser les droits qui risquent d'être bafoués dans chaque situation et définir clairement comment ils seront protégés ;

    e. bien définir la nature de l'approche de régulation de la modération de contenu en précisant le degré d'implication de l'État ;

    f. tirer véritablement les enseignements de l'expérience acquise dans le cadre de systèmes d'autorégulation et de corégulation similaires, afin d'être le plus efficaces possible et de réduire au minimum les conséquences imprévues, et

    g. veiller à ce que toutes les données nécessaires à la transparence soient obtenues et rendues publiques afin que les problèmes existants soient mis en évidence et rectifiés et que le principe de responsabilité soit appliqué.

Définir le problème

15. Les contenus pouvant faire l'objet de modération en raison de politiques publiques sont très divers. Chaque catégorie de contenu représente une problématique distincte en matière de politique publique, qui a ses propres caractéristiques. Les politiques publiques respectives doivent viser à bien distinguer les unes des autres les différentes catégories de contenus illicites et à mettre au point des réponses ciblées, efficaces et proportionnées, adaptées aux particularités du problème concret à traiter[21].

16. Pour réduire au minimum le risque de réponses hâtives, inefficaces, contre-productives ou disproportionnées face à de nouvelles problématiques, les États doivent s'efforcer de définir clairement et de rendre publique une méthodologie qui permette de classer les types de contenus en différentes catégories et d'élaborer des réponses adéquates et respectueuses des droits humains. Les États et les plateformes en ligne devraient faire en sorte qu'il soit possible d'accéder facilement et dans des conditions de transparence à cette méthodologie publique.

17. Pour garantir la proportionnalité, la prévisibilité et l'efficacité des moyens employés, il est essentiel de suivre continuellement l'évolution du problème en considérant une catégorie particulière de contenu illégal, ses moteurs et ses effets sur la société.

18. Il est en particulier primordial que dans le cadre de toute obligation d'autorégulation ou de corégulation visant à lutter contre les crimes graves (par exemple, ceux qui mettent en danger la vie humaine, ou les actes de maltraitance d'enfants), les États soient tenus de prendre toutes les mesures nécessaires pour combattre la composante hors internet de ces crimes. Face à de tels contenus, il ne doit jamais être possible d'adopter une approche d'autorégulation ou de

[21] Par exemple, en présence d'un contenu qui constitue la preuve d'un crime commis hors ligne (comme des actes de maltraitance des enfants), il est nécessaire de prendre des mesures différentes de celles à adopter face à un contenu qui est illégal par son contexte (comme la publication non consensuelle d'images intimes, dite « vengeance pornographique »).

corégulation qui ne fasse pas explicitement référence à la collaboration escomptée avec les services répressifs et les autres autorités publiques concernées[22]. Ces mesures doivent viser à parvenir à une transparence maximale et doivent être très précisément calibrées afin d'éviter les répercussions involontaires sur le respect de la vie privée et d'autres droits humains, et de prévenir tout abus de pouvoir de la part des autorités publiques.

19. Lorsque la modération de contenu est effectuée par des intermédiaires d'internet à des fins commerciales, les États doivent veiller à ce que toute restriction qui en résulterait soit clairement définie, prévisible et imposée d'une manière non discriminatoire qui n'entraîne pas d'ingérence excessive dans l'exercice des droits humains, et à ce qu'il existe des voies de recours adéquates et accessibles.

Assurer la prévisibilité

20. Les restrictions des droits humains doivent être prévisibles, afin de permettre aux individus de réguler leur comportement. Cela porte sur les interdictions de contenu prévues par la loi et les règles en matière de responsabilité imposées aux intermédiaires d'internet par les États. Cela s'applique également à la conception, à la mise à jour et à la mise en œuvre par les intermédiaires d'internet de leurs conditions de service.

21. Les politiques des États en matière de modération de contenu doivent être non discriminatoires et tenir compte des différences considérables de taille et d'échelle qui existent entre les intermédiaires d'internet. Les États devraient éviter de promouvoir des approches qui imposent à ces derniers des obligations disproportionnées, ou qui leur délèguent la prise de décisions au détriment d'approches ayant une légitimité démocratique, notamment en refusant à tout groupe de parties prenantes le droit d'apporter une véritable contribution.

22. De même, en ce qui concerne la modération de contenu effectuée par les intermédiaires d'internet à des fins commerciales, toutes les garanties nécessaires doivent être mises en place en matière de transparence et de procédures pour éviter tout biais discriminatoire ou, si besoin est, mettre en évidence et supprimer de tels biais, en particulier s'ils s'exercent contre des groupes vulnérables.

23. Il incombe à l'État de garantir un cadre juridique prévisible pour toutes les parties concernées. Si les intermédiaires d'internet sont tenus responsables du non-retrait de contenus illégaux, les règles relatives à la « connaissance » de tels contenus qui engage cette responsabilité doivent être claires et proportionnées, tout comme les règles qui interdisent le contenu en question.

24. Pour que la prévisibilité soit garantie, il faut également que la nature et le degré d'implication de l'État dans le mécanisme qui impose la restriction de contenu soient clairs. Les États doivent veiller à ce que, dans tous les cas, les obligations et responsabilités sur le plan juridique, ainsi que les rôles opérationnels et les obligations de rendre compte soient clairement définis.

25. Les États doivent garder à l'esprit que plus il est urgent d'adopter des restrictions, plus il est important d'en assurer l'efficacité et de garantir l'application du principe de responsabilité. Les mécanismes de transparence, de révision et d'ajustement sont essentiels et ne peuvent être relégués au second plan en raison de l'urgence d'une situation.

[22] Par exemple, la loi allemande intitulée *Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (NetzDG)* [loi pour l'amélioration de l'application du droit sur les réseaux sociaux] fait obligation aux intermédiaires d'internet de conserver certaines données connexes lorsqu'ils retirent des contenus, au cas où ces données seraient nécessaires lors d'enquêtes ultérieures.

26. Lorsque les intermédiaires d'internet ont recours à des organisations spécialisées (connues sous le nom de « signaleurs de confiance » ou de « signaleurs prioritaires »)[23] qui servent de filtre permettant d'obtenir des signalements plus fiables des contenus inadmissibles, ces initiatives doivent être soumises à des règles spécifiques en matière de transparence, afin de garantir qu'aucune incitation perverse ou conflit d'intérêt ne soit involontairement créé et que leur degré d'efficacité et de fiabilité demeure constamment élevé. Le recours à des signaleurs de confiance ne devrait pas être obligatoire et les avis de ces derniers ne devraient pas être considérés comme une « connaissance réelle » de l'illégalité d'un contenu. Le statut des « signaleurs de confiance » doit faire l'objet d'évaluations périodiques indépendantes.

27. Si les « signaleurs de confiance » ont pour vocation d'aider à mettre en évidence les contenus à supprimer, il convient de tenir compte du fait que (mis à part un certain nombre d'arrangements informels) il n'existe aucun mécanisme permettant à des groupes de confiance de demander formellement que les erreurs de modération de contenu soient rectifiées. Il y a donc lieu d'étudier la viabilité de systèmes de « dé-signalement de confiance ».

Conformité excessive et discrimination

28. Les États doivent veiller à ce que les mesures d'incitation qui s'appliquent aux intermédiaires d'internet soient équilibrées et éviter que la régulation les incite à imposer des restrictions disproportionnées, ce qui peut, par exemple, se produire lorsque les règles qui définissent la responsabilité des intermédiaires sont soit trop strictes, soit trop vagues. Cela vaut tout particulièrement dans le cas des contenus qui, tout en étant légaux, peuvent être indésirables dans une société démocratique, où il est établi que les droits humains doivent également être respectés[24].

29. Les décisions prises par des êtres humains et par les systèmes technologiques qu'ils créent ne sont pas infaillibles et peuvent être entachées de biais délibérés ou involontaires. Les politiques publiques doivent faire obligation aux intermédiaires d'internet de rendre publiques suffisamment de données pour qu'il soit possible de mener des audits indépendants adéquats permettant de déceler tout aspect discriminatoire ou problématique des décisions prises en matière de restriction de contenu.

30. En outre, les systèmes de modération de contenu et les mécanismes de plainte associés peuvent être utilisés de manière abusive par des acteurs malveillants qui s'en prennent ainsi aux déclarations de groupes auxquels ils s'opposent pour faire en sorte que ces déclarations soient soumises à des restrictions. Cela peut être le fait d'individus au comportement malveillant ou provenir également d'une action concertée (de nombreuses personnes déposant chacune une plainte). Il faut anticiper une telle situation et prendre des mesures correctives adéquates. Les États doivent garder à l'esprit que si des sanctions dissuasives sont essentielles pour lutter contre de tels abus, elles ne constituent pas à elles seules une solution complète.

[23] EuroISPA (2019) « Priority Flagging Partnerships in Practice », disponible à l'adresse www.euroispa.org/wp-content/uploads/Hutty_Schubert_Sanna_Deadman-Priority-Flagging-Partnerships-in-Practice-EuroISPA-2019.pdf, consulté le 28 mai 2020.

[24] Conformément à la jurisprudence de la Cour européenne des droits de l'homme, les idées « qui heurtent, choquent ou inquiètent l'État ou une fraction quelconque de la population » doivent avoir le moyen de s'exprimer – voir, entre autres, *Handyside c. Royaume-Uni*, nº 5493/72, 7 décembre 1976, par. 49.

Droits concernés

31. Toute politique de modération de contenu doit faire l'objet d'une évaluation initiale, puis régulière, qui vise à déterminer si elle a pour effet de restreindre les droits humains. Il convient de prendre des précautions particulières pour veiller au respect de ces droits, les restrictions mises en œuvre devant être pleinement conformes à la Convention européenne des droits de l'homme.

32. Compte tenu des normes existantes du Conseil de l'Europe[25], le droit à un recours effectif signifie que les personnes concernées doivent être informées des raisons précises pour lesquelles leur contenu a été retiré ou leur plainte n'a pas abouti au retrait demandé d'un contenu. Elles doivent avoir droit à un arbitrage accessible. Bien qu'il doive toujours être possible d'accéder aux autorités judiciaires si l'une ou l'autre des parties le souhaite, les États doivent apporter leur appui à la mise en place d'autres mécanismes de règlement des litiges, des solutions d'arbitrage innovantes conçues par plusieurs parties prenantes ou des tribunaux électroniques, selon le cas.

33. Lors de l'octroi de réparations, il doit être tenu compte du fait que le préjudice subi par la partie lésée, et par la société en général, peut ne pas être d'ordre financier et que, dans l'environnement en ligne, la simple publication d'un rectificatif peut ne pas remédier de manière adéquate au préjudice ou à la violation initiale.

34. Il convient également d'accorder une attention particulière aux droits en matière de travail et à la santé mentale de toutes les personnes chargées d'examiner un par un des contenus susceptibles d'être choquants ou perturbants ou d'avoir des répercussions psychologiques. C'est tout particulièrement le cas lorsque les intermédiaires d'internet sous-traitent cette tâche à des tiers, éventuellement établis dans d'autres pays où le droit du travail est différent et moins protecteur.

Nature de l'approche de régulation

35. Les décisions prises par les intermédiaires d'internet en matière de modération de contenu peuvent se situer à différents degrés d'une même échelle, allant de celles qui sont prises en toute indépendance à celles qui le sont sous l'effet direct de pressions exercées par les États, par les médias ou par des campagnes de la société civile par des moyens autres que juridiques ou en conséquence, voulue ou non, des règles qui définissent la responsabilité des intermédiaires face à un vaste ensemble d'infractions. Les États devraient reconnaître et garder à l'esprit que leur degré d'implication dans les décisions que prennent les intermédiaires d'internet influe sur la portée de leurs obligations connexes en matière de droits humains.

36. L'approche corégulatoire de la modération de contenu impliquant de plus importantes obligations pour l'État et lui permettant de garantir un degré plus élevé d'inclusion, de responsabilité et de transparence, elle convient généralement mieux aux contextes faisant intervenir des questions de politiques publiques. Lorsqu'ils pratiquent la corégulation, les États devraient utiliser des formulations et des définitions claires pour ce qui est de la nature de la coopération avec les intermédiaires d'internet, ses objectifs et ses cibles, et les responsabilités et les obligations respectives de toutes les parties concernées.

[25] Voir section 2.5 de la Recommandation CM/Rec (2018)2 du Comité des Ministres aux États membres sur les rôles et les responsabilités des intermédiaires d'internet.

Caractéristiques des approches concluantes et de celles qui ne le sont pas

37. L'étude des caractéristiques d'une autorégulation réussie dans d'autres domaines, qui peuvent également s'appliquer à la corégulation, a permis de mettre en évidence plusieurs facteurs essentiels[26] :

    a. La transparence, notamment en ce qui concerne les objectifs fixés, l'équilibre des pouvoirs et l'indépendance ;

    b. Des repères et des cibles objectifs, clairement définis et vérifiés de manière indépendante ;

    c. La publication d'informations et la réalisation d'évaluations et d'audits indépendants obligatoires portant sur le respect des codes et les progrès accomplis dans la réalisation des objectifs, et l'imposition de sanctions adéquates en cas de non-respect[27] ;

    d. Une surveillance indépendante continuelle.

38. Il convient d'examiner régulièrement, au vu de l'expérience acquise, lesquelles des approches d'autorégulation ou de corégulation ont été entièrement ou partiellement couronnées de succès et lesquelles ne l'ont pas été, afin d'améliorer en permanence l'efficacité de ces approches en matière de modération de contenu et leur compatibilité avec les droits humains.

Transparence

39. La souplesse d'utilisation de l'autorégulation et de la corégulation est leur principal avantage, qui est particulièrement appréciable dans un environnement en ligne en constante évolution, où il faut apporter rapidement des réponses adéquates aux problèmes qui se font jour. Les États et les intermédiaires d'internet doivent reconnaître et garder à l'esprit qu'en l'absence d'une véritable transparence, la société se prive de cet avantage de l'autorégulation et de la corégulation, tout en devant malgré tout subir la diminution de la responsabilité et de la légitimité démocratique qui constituent leur aspect négatif[28].

40. Pour que la modération de contenu respecte les droits humains et réponde à ses objectifs, la transparence est essentielle. Elle est nécessaire pour :

    – définir clairement la nature de la modération du contenu (par exemple, si sa mise en œuvre relève de l'autorégulation ou de la corégulation et pourquoi, et si elle est effectuée directement ou sous l'effet de mesures d'incitation intégrées) ;

    – bien cerner le problème à régler et les objectifs fixés à cet égard ;

    – définir clairement les droits qui risquent d'être restreints ;

    – présenter clairement tout processus entièrement ou partiellement automatisé éventuellement utilisé pour prendre des décisions en matière de modération de contenu ;

---

[26] Sharma, L. L., Teret, S. P., & Brownell, K. D. (2010), « The food industry and self-regulation: standards to promote success and to avoid public health failures », *American journal of public health*, 100(2), 240–246, disponible à l'adresse https://doi.org/10.2105/AJPH.2009.160960, consulté le 8 octobre 2020.

[27] Commission européenne (2016), étude sur « L'efficacité de l'autorégulation et de la corégulation dans le contexte de la mise en œuvre de la directive "Services de médias audiovisuels" », disponible à l'adresse https://ec.europa.eu/digital-single-market/en/news/audiovisual-and-media-services-directive-self-and-co-regulation-study, consulté le 6 mai 2021.

[28] En l'absence de véritable transparence, il n'est pas possible de cerner et d'évaluer les changements qui surviennent ni de procéder aux ajustements correspondants. Voir également la section 2.2 de la Recommandation CM/Rec (2018)2 du Comité des Ministres aux États membres sur les rôles et les responsabilités des intermédiaires d'internet.

– définir clairement, le cas échéant, quels types de contenus ou de comportements non illégaux ne sont pas autorisés dans le cadre des services d'un intermédiaire d'internet ;

– faire en sorte que la solution la moins restrictive soit retenue en cas de limitation de droits ;

– mettre en évidence et éliminer les erreurs qui conduisent au retrait de contenus légitimes ou au maintien en ligne de contenus illégitimes.

41. Afin de garantir une transparence et des audits adéquats, il est primordial que les intermédiaires d'internet conservent, dans le plein respect de la législation et des principes relatifs à la protection des données, les contenus qui ont été soumis à des restrictions et les raisons pour lesquelles ils l'ont été.

42. Les États doivent savoir que la rapidité et la quantité de contenus supprimés ne témoignent pas nécessairement de l'efficacité des mesures prises. L'efficacité que les indicateurs devraient viser à évaluer se manifeste par les progrès accomplis dans la réalisation d'objectifs concrets de politiques publiques.

43. De même, l'un des effets positifs de la transparence est de permettre d'effectuer un suivi des problèmes entre les différents intermédiaires d'internet et dans le temps. Si les méthodologies et les formats des rapport des intermédiaires d'internet ne permettent pas un tel suivi, les décideurs politiques ne pourront pas tirer parti de cet avantage.

    Pour parvenir à une transparence entière et réelle, les données devraient être fournies avec un degré de précision maximal et selon la méthodologie la plus cohérente possible, pour des types d'intermédiaires d'internet similaires et dans le temps, afin de permettre de bien analyser et évaluer les méthodes de modération de contenu appliquées, et elles devraient être mises à la disposition du public dans des termes clairs.

Exposé des motifs

à la note d'orientation sur les meilleures pratiques en matière de cadres juridiques et procéduraux efficaces pour les mécanismes d'autorégulation et de co-régulation de la modération de contenu

Concepts clés

L'objectif de cette section est de présenter les concepts clés qui seront utilisés tout au long de l'exposé des motifs et dans la note d'orientation.[29]

Censure : Restriction de l'utilisation de certaines images, mots, opinions ou idéologies. Le mot est utilisé ici dans le sens juridique anglais neutre.[30]

Curation du contenu : Processus consistant à décider quel contenu doit être présenté aux utilisateurs (en termes de fréquence, d'ordre, de priorité, etc.), sur la base du modèle commercial et de la conception de la plate-forme.

Modération du contenu : Processus par lequel une entreprise qui héberge un contenu en ligne évalue la [il]légalité ou la compatibilité avec les conditions de service d'un contenu tiers, afin de décider si certains contenus mis en ligne, ou dont on a tenté la mise en ligne, doivent être rétrogradés (c'est-à-dire laissés en ligne mais rendus moins accessibles), marqués comme potentiellement inappropriés ou incorrects, démonétisés,[31] non sanctionnés ou retirés, pour tout ou partie du public, par le service sur lequel ils ont été mis en ligne.

La **corégulation** : Mesures prises de manière proactive par des entreprises ou des secteurs, en coopération avec les États ou sous leur supervision, pour

- démontrer le respect d'une obligation légale ou d'accords ou de codes non contraignants ou,
- réglementer leurs activités, à la suite d'une négociation ou d'une coopération avec les États, ou à la suite d'un encouragement des États. Dans le cadre de l'UE, il peut s'agir d'un "mécanisme par lequel un acte législatif communautaire confie la réalisation des objectifs définis par l'autorité législative à des parties reconnues dans le domaine".[32]

---

[29] Une discussion approfondie sur ces concepts et d'autres concepts connexes est, à l'heure où nous écrivons ces lignes, en cours au sein de la Coalition du Forum sur la gouvernance de l'Internet sur la responsabilité des plateformes. Pour plus d'informations, voir www.intgovforum.org/multilingual/content/glossary-on-platform-law-and-policy-terms , consulté le 15 janvier 2020.

[30] Bien que ce mot soit souvent utilisé dans un sens péjoratif, il est utilisé ici dans un sens juridique neutre, comme il est courant dans les pays anglophones. Par exemple, le mandat de l'Irish Board of Film Classification est établi par la loi sur la censure des films de 1923 et le directeur de la classification des films est officiellement le "censeur officiel". Voir www.ifco.ie/en/ifco/pages/legislation , consulté le 6 janvier 2020. L'organisme britannique équivalent était appelé "British Board of Film Censors" jusqu'en 1984.

[31] Certaines plateformes rémunèrent les utilisateurs pour le contenu téléchargé (ce qui permet à l'utilisateur de "monétiser" son contenu) et peuvent supprimer ces paiements, dans certaines circonstances. Pour des informations sur la politique de Google/YouTube, voir https://support.google.com/youtube/answer/6162278 , consulté le 13 octobre 2020.

[32] Accord interinstitutionnel "Mieux légiférer" de 2003, 2003/C 321/01, paragraphe 18. Voir https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.C_.2003.321.01.0001.01.ENG , consulté le 10 juin 2020. Dans ce contexte, ce document indique clairement que les mécanismes de corégulation et d'autorégulation "ne seront pas

Un modèle de corégulation qui fournit une base juridique à un organisme d'autorégulation et manifestement indépendant peut offrir une approche qui respecte les normes internationales en matière de liberté d'expression.

Bien que la directive européenne sur les services de médias audiovisuels (SMAV) ne définisse pas la corégulation, elle fournit une description de ce qu'elle est censée être.[33]

Règlement : Une règle/obligation légale imposée par l'État ou l'acte de contrôler quelque chose.

L'**autorégulation** : Mesures volontaires prises par des entreprises ou des secteurs sans l'encouragement du gouvernement. Un considérant de la directive SMA de l'UE décrit l'autorégulation comme suit : "L'autorégulation constitue un type d'initiative volontaire qui permet aux opérateurs économiques, aux partenaires sociaux, aux organisations non gouvernementales et aux associations d'adopter des lignes directrices communes entre eux et **pour eux-mêmes**. Ils sont chargés d'élaborer, de contrôler et de faire respecter ces lignes directrices".[34] (soulignement ajouté)

En outre, dans la pratique, les termes "autorégulation" et "corégulation" ont été utilisés avec un degré de chevauchement considérable et peuvent être subdivisés en plusieurs sous-catégories.[35]

La modération du contenu soulève des défis particuliers en ce qui concerne le rôle et la responsabilité des États.[36] Ces préoccupations découlent du fait qu'il s'agit d'une tâche généralement mise en œuvre par des parties privées, souvent pour atteindre des objectifs de politique publique. Par conséquent, ces explications de concepts soulignent le fait que les États sont les principaux responsables des droits de l'homme.

I Introduction

L'internet nous a donné de nouvelles possibilités fantastiques de parler, d'être entendus et de nous organiser. En effet, il a créé une multitude de nouvelles possibilités d'exercer nos droits humains, notamment nos droits à la liberté d'expression, à la liberté de réunion, à la liberté de pensée et de religion et autres. Toutefois, il n'est pas surprenant qu'il crée également des possibilités de diffusion de contenus ou de comportements illégaux ou potentiellement préjudiciables. Les services en ligne (tels que les plateformes de médias sociaux, où les gens publient des messages, des articles, des photos, etc.), disposent d'un processus de suppression, de rétrogradation ou d'autres moyens de décourager la diffusion de contenus illégaux ou indésirables, appelé "modération de contenu".

---

applicables lorsque des droits fondamentaux ou des options politiques importantes sont en jeu" (paragraphe 17). Cet instrument n'est plus en vigueur. Son successeur ne mentionne pas la corégulation et l'autorégulation.

[33] "La corégulation fournit, dans sa forme minimale, un lien juridique entre l'autorégulation et le législateur national conformément aux traditions juridiques des États membres. Dans la corégulation, le rôle réglementaire est partagé entre les parties prenantes et le gouvernement ou les autorités ou organismes réglementaires nationaux : Directive (UE) 2018/1808 du Parlement européen et du Conseil du 14 novembre 2018 modifiant la directive 2010/13/UE visant à la coordination de certaines dispositions législatives, réglementaires et administratives des États membres relatives à la fourniture de services de médias audiovisuels (directive «Services de médias audiovisuels»), compte tenu de l'évolution des réalités du marché, considérant 14. Voir https://eur-lex.europa.eu/eli/dir/2018/1808/oj , consulté le 19 février 2021.

[34] Ibid.

[35] Marsden C.T. (2012) *Internet co-regulation and constitutionalism : Vers un contrôle juridictionnel européen International Review of Law Computers & Technology*, 26(2):211-228, www.researchgate.net/publication/254294662_Internet_co-regulation_and_constitutionalism_Towards_European_judicial_review , consulté le 10 juin 2020.

[36] Husovec Martin (2021) "Surblocage : When is the EU legislator responsible", https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3784149 , consulté le 17 février 2021.

Un grand nombre d'excellentes recherches et analyses ont déjà été effectuées sur les aspects opérationnels de la modération de contenu, par des organisations internationales comme les Nations Unies, des organisations non gouvernementales, individuellement et collectivement, et des universitaires. L'objectif de la note d'orientation n'est pas de reproduire ou même de cataloguer de manière exhaustive cet impressionnant corpus de travaux. Il convient de noter notamment les travaux

- des anciens rapporteurs spéciaux des Nations unies, Frank LaRue[37] et David Kaye,[38]
- du Comité des Ministres du Conseil de l'Europe,[39] y compris les recherches qu'il a commandées.[40]
- d'innombrables instituts universitaires tels que l'Institut de droit de l'information d'Amsterdam,[41] le FGV Direito Rio[42] et le Centre pour l'Internet et la société de l'université de Stanford.[43]
- ou encore les travaux coordonnés et/ou réalisés par des ONG telles que AccessNow,[44] Article,[45] Electronic Frontier Foundation,[46] epicenter.works,[47] European Digital Rights[48] et Meedan.[49]

La liste des travaux remarquables dans ce domaine est bien trop longue pour en citer tous les exemples.

Les questions relatives à la modération de contenu ont également été étudiées par le projet Internet & Juridiction, de manière générale en relation avec les questions de juridiction, et tout récemment dans le Rapport de situation globale 2019 et[50] sous la forme de deux séries de recommandations spécifiques et granulaires sur les "Cartographies des éléments clés de la modération de contenu".[51]

---

[37] Conseil des droits de l'homme des Nations unies, Rapport du rapporteur spécial sur la promotion et la protection du droit à la liberté d'opinion et d'expression, Frank La Rue, 2017, www.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf , consulté le 23 septembre 2020.

[38] David Kaye (2019) "A New Constitution for Content Moderation", medium.com, https://onezero.medium.com/a-new-constitution-for-content-moderation-6249af611bdf , consulté le 23 septembre 2020.

[39] Recommandation CM/Rec(2020)1 du Comité des Ministres aux États membres sur les implications des systèmes algorithmiques pour les droits de l'homme et Recommandation CM/Rec(2018)2 du Comité des Ministres aux Etats membres sur les rôles et responsabilités des intermédiaires d'internet.

[40] Comme Brown Alexander (2020) "Models of Governance of Hate Speech Online", https://rm.coe.int/models-of-governance-of-online-hate-speech/16809e671d , consulté le 15 janvier 2021.

[41] Voir https://www.ivir.nl/publications/technology-and-law/.

[42] Belli Luca, Zingales Nicolo (2017) *Platform Regulations : How platforms are regulated and how they regulate us*, ISBN 978-85-9597-014-4, https://direitorio.fgv.br/publicacoes/platform-regulations-how-platforms-are-regulated-and-how-they-regulate-us , consulté le 19 février 2021.

[43] Voir https://cyberlaw.stanford.edu/focus-areas/intermediary-liability.

[44] AccessNow (2020) "26 recommandations sur la gouvernance du contenu", www.accessnow.org/cms/assets/uploads/2020/03/Recommendations-On-Content-Governance-digital.pdf , consulté le 23 septembre 2020.

[45] Article 19 (2018) "Droits d'évitement : Regulating Speech by Contract", www.article19.org/wp-content/uploads/2018/06/Regulating-speech-by-contract-WEB-v2.pdf , consulté le 4 janvier 2020.

[46] Voir santaclaraprinciples.org , consulté le 23 septembre 2020.

[47] Voir platformregulation.eu , consulté le 23 septembre 2020.

[48] Voir https://edri.org/?s=content+moderation , consulté le 23 septembre 2020.

[49] Meedan (2018) "Content Moderation Toolkit", https://meedan.com/reports/content-moderation-toolkit/ , consulté le 19 février 2021.

[50] Dan Jerker B. Svantesson, "Internet and Jurisdiction Global Situation Report 2019", www.internetjurisdiction.net/uploads/pdfs/GSR2019/Internet-Jurisdiction-Global-Status-Report-2019_web.pdf , consulté le 28 septembre 2020.

[51] Projet Internet & Juridiction (2020) "I&J Outcomes : Mappings of Key Elements of Content Moderation", www.internetjurisdiction.net/news/i-j-outcomes-mappings-of-key-elements-of-content-moderation , consulté le 28 septembre 2020.

Au lieu de reformuler ou de réimaginer l'ensemble des travaux, la note d'orientation et le présent exposé des motifs prennent du recul et examinent le cadre de la modération de contenu, afin d'établir des lignes directrices de haut niveau à l'intention des États sur l'élaboration d'approches de la modération de contenu qui soient à la fois compatibles avec les droits de l'homme et qui permettent d'atteindre leurs objectifs de politique publique, ainsi que pour guider les entreprises privées.

La note d'orientation aborde des questions plus larges, notamment, comment mieux comprendre la nature des problèmes spécifiques que la modération de contenu cherche à résoudre, comment garantir une responsabilité appropriée en cas de restrictions des droits de l'homme, ainsi que les concepts d'autorégulation et de corégulation, et les caractéristiques des outils de transparence qui sont fondamentaux pour garantir que les objectifs soient fixés et atteints.

Comprendre les problèmes

La modération de contenu est un outil utilisé pour traiter une grande variété de problèmes différents. C'est un élément de la lutte contre la criminalité grave en ligne, contre d'autres infractions en ligne, contre les contenus qui peuvent être préjudiciables à certains publics et contre les contenus qui peuvent être problématiques pour le modèle commercial des entreprises en ligne (contenus hors sujet sur une plateforme spécialisée, par exemple).

Une fois que l'on se penche sur les particularités de ces défis, on constate qu'une solution unique peut souvent être possible, mais qu'elle est rarement souhaitable. Les conséquences de la simple suppression (ou non) d'un post hors sujet dans un forum spécialisé sont radicalement différentes de celles d'une plateforme qui supprime simplement une vidéo d'un crime grave en cours de réalisation, par exemple.

Quel que soit le problème traité par la modération de contenu, la suppression d'un message en ligne est une limitation de la liberté d'expression d'un utilisateur, et doit donc être effectuée d'une manière prévisible, légitime, nécessaire et proportionnée.

Les États ne doivent pas non plus supposer que les intermédiaires d'internet sont les mieux placés pour prendre des décisions sur la légalité ou l'illégalité du contenu ou qu'ils sont neutres lorsqu'ils prennent de telles décisions. Lorsqu'il s'agit de décider de laisser un contenu en ligne, une société privée serait mal placée pour trouver un équilibre entre les droits du plaignant et ceux de la personne qui télécharge le contenu. Elle est encore moins apte à équilibrer les droits lorsque ses propres intérêts sont en jeu.[52] Le cours de l'action de Twitter a chuté de plus de 10 % après la suspension définitive du compte du président américain de l'époque, Donald Trump, en raison de l'impact attendu sur l'engagement avec le contenu sur la plateforme.[53]

Qui est responsable ?

Lorsque nous examinons la responsabilité de ces restrictions, nous devons tenir compte du fait qu'aucune modération du contenu n'est parfaite et qu'elles peuvent être imposées en tant que

[52] Pour une analyse détaillée des questions relatives à l'"équilibre des droits", voir, Douek Evelyn (2021) "Governing Online Speech : From 'posts-as-trumps' to proportionality and probability", *Columbia Law Review*, Vol. 121, No. 1, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3679607 , consulté le 15 janvier 2021.

[53] "Twitter perd 5 milliards de dollars en valeur marchande après l'interdiction définitive de la plateforme Trump", *Business Insider*, 11 janvier 2020, https://markets.businessinsider.com/news/stocks/twitter-stock-price-president-donald-trump-permanently-banned-tweeting-2021-1-1029946778 , consulté le 11 janvier 2020.

décisions privées d'intermédiaires d'internet, décisions directement imputables à la réglementation de l'État ou un mélange des deux.

"L'ampleur insondable de la parole en ligne fait que l'application des règles n'est jamais qu'une question de probabilité : la modération du contenu impliquera toujours des erreurs, et la question pertinente est donc de savoir quels taux d'erreur sont raisonnables et quels types d'erreurs doivent être préférés".[54] Il faut donc clarifier la responsabilité de la fixation de ces objectifs, des restrictions qui en découlent et de la transparence permettant de contrôler utilement ces décisions.

Traditionnellement, les intermédiaires d'internet ont préféré ne pas publier de données significatives sur leurs propres décisions, bien qu'ils aient été transparents sur les décisions prises par les États et les autres qui leur ont été imposées. Cela reflète un intérêt personnel à ne pas attirer l'attention sur leurs décisions. Comme c'est généralement le cas dans ce type de politique, on ne peut pas s'attendre à ce que les intermédiaires d'internet agissent volontairement contre leur propre intérêt et, par conséquent, des obligations légales spécifiques en matière de transparence et de méthodologie sont nécessaires.

L' "autorégulation" est comprise différemment selon les contextes

Nous sommes habitués à l'autorégulation dans l'environnement médiatique traditionnel. Les entreprises de médias établissent leurs règles pour la qualité de leur production et de leurs décisions éditoriales, comme un mécanisme permettant de maintenir leur indépendance et leur qualité. Elles s'autorégulent littéralement.

La situation est beaucoup plus compliquée en ce qui concerne la modération du contenu. Il peut s'agir d'un processus entièrement interne concernant, par exemple, la rapidité de traitement des plaintes relatives au contenu, ce qui correspond parfaitement à la notion d'"autorégulation". Cependant, les décisions de retrait d'un contenu d'utilisateur sont une régulation de la parole des utilisateurs, donc non pas littéralement une régulation de "soi", mais une régulation des utilisateurs et, indirectement, une régulation de ceux qui auraient autrement rencontré ce contenu. Cela soulève des considérations différentes, mais aussi très sérieuses, pour la démocratie et les droits de l'homme.

En outre, les initiatives d'autorégulation peuvent être entreprises en coopération avec les gouvernements et/ou avec leur encouragement, ce qui en fait une entreprise de coopération entre l'industrie et les gouvernements, ou plus proche de la "corégulation". La mesure dans laquelle la modération du contenu engage la responsabilité de l'État en matière de droits de l'homme dépend fortement de la place de l'activité spécifique dans le continuum entre l'autorégulation et la corégulation. Les États peuvent également être tenus responsables s'ils ne prennent pas de mesures ou s'il n'existe pas de réglementation pour prévenir/réparer les violations.

L'ampleur de l'activité est également entièrement différente par rapport à l'autorégulation traditionnelle des médias. Une chaîne de télévision diffusant 24 heures sur 24 génère chaque jour une quantité spécifique et prévisible de vidéos, tandis que 24 heures de vidéos sont téléchargées sur YouTube toutes les trois secondes.[55]

[54] Douek Evelyn (2021), op cit., p. 1.

[55] « Heures de vidéo téléchargées sur YouTube chaque minute à partir de mai 2019 », *Statista*, www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/ , consulté le 13 octobre 2020.

Succès ou échec ?

La modération de contenu est un outil. Elle est utilisée pour atteindre un objectif ou, comme nous l'avons vu, toute une série d'objectifs en utilisant toute une série de structures, telles que l'autorégulation et la corégulation pour atteindre ces objectifs. Mais que faire si cela ne fonctionne pas ? Et si les objectifs ne sont jamais clairement définis ? Et si cela fonctionne, mais que le coût est trop élevé ? Et si elle commence à fonctionner et qu'elle cesse ensuite de fonctionner ? Et si elle n'a pas d'objectif clair ? Ces questions, et les réponses à celles-ci, ne peuvent être abordées que si elles sont clairement formulées et si les données sont collectées pour permettre d'y répondre.

II Énoncé du problème

Cette section examine les questions clés soulevées par la modération du contenu.

La section commence par examiner le fait qu'un large éventail de types de contenus différents peut être soumis à une réglementation des contenus (allant des interdictions légales explicites à la modération des contenus légaux), avec des implications variables pour les droits de l'homme.

La section examine ensuite comment les règles internes des sociétés Internet sont élaborées et appliquées, en accordant une attention particulière à la normalisation et à l'application des règles, ainsi qu'aux technologies utilisées pour faire respecter ces règles.

Enfin, la section examine l'équilibre complexe des rôles et des responsabilités des États et des entreprises privées, dans un environnement où les parties privées imposent souvent des restrictions à la suite de pressions directes ou indirectes du gouvernement et/ou en raison de la responsabilité pour défaut d'application des lois, qui sont parfois peu claires. Elle effectue cette analyse sous cinq angles :

- les implications de la privatisation des politiques de l'État sur la restriction du contenu
- le défi consistant à assurer la prévisibilité des restrictions de contenu de toutes sortes qui sont mises en œuvre par l'application de conditions de service privées ;
- le manque persistant de clarté concernant l'équilibre entre les commandes privées et la réglementation publique en ce qui concerne, par exemple, l'application de lois ou de procédures peu claires ou mal formulées qui ont un impact direct ou indirect sur la modération des contenus ;
- les restrictions imposées pour des raisons d'ordre public sans que celles-ci ne soient prescrites par la loi ;
- la nécessité de choisir les mesures avec soin, afin d'éviter les mauvaises pratiques (telles qu'une attention excessive portée à la vitesse ou au volume du contenu retiré).

1. Prévalence des contenus/comportements importuns ou abusifs

Il est essentiel de reconnaître que les différents sujets de la modération de contenu sont des problèmes fondamentalement différents et, par conséquent, que des solutions universelles, toutes faites, peuvent ne pas être appropriées. Le chapitre IV détaille six grandes catégories de contenus illégaux ou potentiellement problématiques, allant des contenus illégaux partout aux contenus légaux mais potentiellement préjudiciables. Il convient de noter que, à l'exception partielle des images d'abus d'enfants, il n'existe guère de législation harmonisée au niveau international sur ces types de contenu. Cela soulève des problèmes juridictionnels importants lorsque la personne qui met à disposition du

contenu sur internet se trouve dans un pays, qu'une autre personne télécharge ce contenu dans un deuxième pays et que le fournisseur de services se trouve dans un pays tiers.[56] Cela accroît l'insécurité juridique pour les intermédiaires d'internet et, par conséquent, les risques, par exemple, de surblocage, de mise en œuvre extraterritoriale des lois nationales, etc.[57]

Il est donc important pour les États de concevoir une méthodologie structurée pour répondre rapidement, proportionnellement et efficacement à ces problèmes importants. Alors qu'elle est essentielle, une telle approche, à l'heure actuelle, fait défaut. Par exemple, les réponses politiques à la propagation des théories de conspiration sur des sujets aussi divers que les communications mobiles 5G et les vaccins ont jusqu'à présent été généralement lentes et inadéquates.[58] Ces contenus peuvent avoir des conséquences importantes dans le monde réel, allant de l'incendie des pylônes de communication mobile à des épidémies de maladies qui étaient auparavant sous contrôle. Un autre exemple est la montée redoutée des "contrefaçons profondes". Cette technologie permet de recréer de manière convaincante des vidéos existantes, en remplaçant un élément visuel ou audio (par exemple une personne) dans la vidéo par quelqu'un/quelque chose d'autre. C'est un exemple de contenu qui ne fait pas nécessairement partie d'une infraction plus large et où le contenu lui-même n'est pas nécessairement illégal. Il est préférable de mettre en place des stratégies qui peuvent être utilisées quand de tels phénomènes sont susceptibles de devenir un problème, plutôt que de réagir lorsqu'ils le deviennent.

2. Absence de normalisation des restrictions en matière de contenu

Dans les régimes d'autorégulation et de corégulation, les restrictions sont généralement imposées sur la base des règles internes de l'intermédiaires d'internet. Le nom et le nombre de ces règles internes varient d'une entreprise à l'autre.[59]

Il est symptomatique de la complexité des questions en jeu que l'on soit confrontés à la fois à un manque de normalisation (concernant la manière dont les entreprises individuelles interprètent et appliquent leurs règles internes) et à une trop grande normalisation unilatérale de la part des plus grands fournisseurs.

Nous souffrons d'un manque de normalisation dans la mesure où la signification des mots dans les termes de service des intermédiaires de l'internet est souvent peu claire. Par ailleurs, il est encore plus problématique que les termes utilisés sont également sujets à réinterprétation.

À titre d'exemple, Propublica a constaté que les modérateurs de contenu de Facebook arrivaient à des conclusions différentes sur la nécessité de supprimer des contenus largement similaires et ne respectaient pas toujours les directives de l'entreprise sur la manière de traiter le contenu.[60] En

[56] Pour une introduction à ce sujet, voir : Smith Graham (2018) "Peaceful coexistence, jurisdiction and the internet", www.cyberleagle.com/2018/02/peaceful-coexistence-jurisdiction-and.html , consulté le 1er septembre 2020.

[57] Voir Dan Jerker B. Svandesson (2019), op cit.

[58] Voir, par exemple, EU Disinfo Lab, "Covid-19 and 5G : A Case Study of Platforms' Content Moderation of Conspiracy Theories", 14 avril 2020, www.disinfo.eu/publications/coronavirus-and-5g-a-case-study-of-platforms-content-moderation-of-conspiracy-theories, consulté le 4 janvier 2021.

[59] Facebook, par exemple, a des "conditions de service" (www.facebook.com/legal/terms) et des "normes communautaires" (https://facebook.com/communitystandards) pour ses utilisateurs de services de médias sociaux, tandis que Twitter a des "règles de Twitter". (https://help.twitter.com/en/rules-and-policies/twitter-rules), consultés le 8 octobre 2020.

[60] Tobin Ariana, Varner Madeleine et Angwin Julia, "Facebook's Uneven Enforcement of Hate Speech Rules Allows Vile Posts to Stay Up", *Propublica*, 28 décembre 2017, www.propublica.org/article/facebook-enforcement-hate-speech-rules-mistakes , consulté le 5 mai 2020.

substance, des contenus largement similaires à la fois sont, et ne sont pas, interdits par Facebook, et ses mécanismes d'application de ses propres règles internes ne sont pas toujours respectés.

Ce problème est exacerbé par le fait que les « conditions de service » sont rédigées de manière vague, peut-être délibérément, afin de donner aux intermédiaires d'internet un maximum de souplesse pour agir s'ils sont mis sous pression par les États, par exemple.[61] Une illustration particulièrement claire de la signification flexible des accords de Facebook avec ses utilisateurs se trouve dans son expérience inattendue de 2012 visant à établir si la société pouvait manipuler l'état d'esprit d'environ 70 000 de ses utilisateurs. Face aux critiques des médias, Facebook a expliqué que les utilisateurs s'étaient inscrits pour être « recherchés" de cette manière parce que "quand quelqu'un s'inscrit sur Facebook, nous avons toujours demandé la permission d'utiliser ses informations pour fournir et améliorer les services que nous offrons. Suggérer que notre entreprise a effectué de telles recherches sans autorisation est une pure fiction."[62] Mais curieusement, l'entreprise a ajouté le terme "recherche" à ses conditions de service quelques mois plus tard, alors qu'elle avait déjà déclaré qu'il s'agissait clairement d'une utilisation possible des données personnelles.

Nous souffrons d'une trop grande standardisation dans la mesure où les aspects clés des conditions de service et l'utilisation d'outils spécifiques de filtrage par les plus grands intermédiaires d'internet créent de facto des normes mondiales ayant un impact potentiellement important sur les droits de l'homme, sans que la prise de décision soit multipartite ou légitimée démocratiquement. Ainsi, alors que les conditions de service sont souvent vagues et imprévisibles, les restrictions de contenu qui en découlent sont harmonisées/normalisées par les intermédiaires d'internet mondiaux. Il s'agit bien entendu d'une arme à double tranchant, plus ils s'accordent une marge d'arbitraire, plus cet arbitraire peut être exploité par des pressions extérieures.

Les règles et les technologies utilisées pour les mettre en œuvre à l'échelle deviennent plus homogènes, comme l'explique Evelyn Douek dans son essai "The Rise of the Content Cartels".[63] Mme Douek décrit les processus qui ont conduit un très petit nombre de grandes sociétés Internet à fixer la norme pour ce qui est autorisé en ligne au niveau mondial, ainsi que les technologies utilisées pour mettre en œuvre cette norme, au détriment de la transparence et de la responsabilité. Cela se fait par le biais d'initiatives d'autorégulation nominales telles que le Forum mondial de l'Internet pour la lutte contre le terrorisme (GIFCT).[64] Ces initiatives commencent normalement par la participation des principales entreprises mondiales et de certains États, qui fixent les règles et définissent les technologies et les listes de filtrage à mettre en œuvre. Elles sont ensuite rejointes, souvent sous la pression des gouvernements, par de plus petites entreprises, qui n'ont guère de contrôle sur les restrictions qu'elles imposent "volontairement" ou qui n'y participent pas.[65]

---

[61] Article 4(1)m du règlement 2016/794 du Parlement européen et du Conseil, 11 mai 2016, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0794 , consulté le 7 mai 2020.

[62] Hern Alex, "Facebook T&Cs introduced 'research' policy months after emotion study", *The Guardian*, 1 juillet 2014, www.theguardian.com/technology/2014/jul/01/facebook-data-policy-research-emotion-study , consulté le 5 mai 2020.

[63] Doudek Evelyn (2020) "The Rise of Content Cartels", Knight First Amendment Institute, https://knightcolumbia.org/content/the-rise-of-content-cartels , consulté le 5 mai 2020.

[64] Voir gifct.org , consulté le 20 mai 2020.

[65] La Recommandation de la Commission européenne (2018)1177 sur des mesures visant à lutter efficacement contre les contenus illicites en ligne (mars 2018) souligne la nécessité de s'engager avec les petits opérateurs pour les aider à déployer des technologies de restriction des contenus. Par exemple, le considérant 37 qui recommande le déploiement de mesures proactives en matière de contenu aux entreprises qui n'ont pas la taille nécessaire pour les exploiter. "Ces efforts de coopération sont particulièrement importants pour permettre aux fournisseurs de services d'hébergement qui, en raison de leur taille ou de l'échelle à laquelle ils opèrent, disposent de ressources et de compétences limitées, de répondre efficacement et de manière urgente aux demandes de renvoi et de prendre des mesures proactives, comme cela est

Bien qu'elle ne respecte pas les principes du multipartisme,[66] une telle normalisation ne porte pas nécessairement atteinte aux droits de l'homme. Pour garantir la protection des droits de l'homme, les États doivent s'assurer que trois conditions essentielles sont remplies : a) il existe un accord clair sur l'illégalité du contenu qui est restreint, en pleine conformité avec la législation sur les droits de l'homme, b) il y a une transparence concernant l'engagement des autorités chargées de l'application de la loi dans les cas où des preuves de tels crimes sont détectées et, c) il y a une transparence rigoureuse concernant le contenu qui est supprimé, dont les critères sont décrits dans la section V ci-dessous. Afin d'avoir une restriction efficace et légale des contenus illégaux, nous avons besoin de règles claires, avec un maximum de transparence. Lorsque les règles ne sont pas claires et/ou font l'objet de modifications régulières et mal communiquées, qu'elles sont conçues de manière opaque et imposées arbitrairement, ni la portée des règles ni les sanctions en cas de violation ne sont connues. En outre, l'impact sur le(s) crime(s) visé(s) n'est pas clair.

Ainsi, à un niveau, nous avons des conditions de service peu claires qui ne sont pas appliquées de manière cohérente par les différents intermédiaires d'internet et, à un autre niveau, nous avons des géants mondiaux de l'internet qui fixent des normes à la fois pour ce qui est du contenu filtré et pour la manière dont il l'est.

En outre, en ce qui concerne les contenus qui se trouvent dans une zone grise, tels que ceux qui peuvent avoir "l'intention" d'"inciter au terrorisme", la loi ne peut guère aider un intermédiaire d'internet à pouvoir évaluer l'illégalité, rôle qu'une partie intéressée est de toute façon mal adaptée à remplir. Lorsque ce contenu est retiré, il n'y a actuellement aucune transparence concernant, par exemple, l'implication des autorités chargées de l'application de la loi ou l'évaluation indépendante ex post de l'illégalité du contenu restreint. Lorsque les technologies qui peuvent être utilisées sont intrusives ou disproportionnées, comme le blocage de certains mots, phrases et images, ou le recours à des technologies pour « deviner » l'intention derrière des mots, phrases ou images, la prévisibilité des restrictions de contenu devient encore plus faible.

> "*Les « cartels de contenu » actuels, ainsi que les futurs « cartels » vers lesquels nous nous dirigeons probablement, permettent aux participants de blanchir des décisions difficiles par des processus opaques afin de les faire paraître plus légitimes qu'elles ne le sont réellement et n'atténuent pas la menace d'une poignée d'acteurs détenant trop de pouvoir sur la sphère publique*". [67]

Toutes les mesures prises pour restreindre le contenu sont des restrictions de la liberté d'expression. Les États ont des obligations positives et négatives de veiller à ce que les restrictions imposées par les régimes de réglementation, d'autorégulation et de corégulation soient juridiquement acceptables. Pour être acceptables, elles doivent être prévisibles, légitimes, nécessaires et proportionnées. Le manque de clarté et l'absence de contrôle indépendant signifient que ces obligations ne sont pas respectées actuellement.

Cela soulève des questions cruciales que les décideurs politiques doivent prendre en compte :

recommandé. (https://ec.europa.eu/digital-single-market/en/news/commission-recommendation-measures-effectively-tackle-illegal-content-online , consulté le 14 octobre 2020).

[66] Le multipartenariat est un principe de gouvernance de l'internet largement promu, dont l'objectif est de garantir que les intérêts de toutes les parties prenantes sont dûment pris en considération dans la conception et la mise en œuvre de solutions à des problèmes communs ou pour atteindre des objectifs communs.

[67] Doudek Evelyn (2020) "The Rise of the Content Cartels", op cit.

- Premièrement, les décideurs politiques doivent tenir compte du fait que la modération de contenu s'attaque à une grande variété de types de contenus indésirables et qu'une approche unique a peu de chances d'être efficace ou proportionnée.
- Deuxièmement, les restrictions imposées dans le cadre de la modération de contenu sont généralement basées sur les règles internes des intermédiaires d'internet, qui sont souvent peu claires et mises en œuvre de manière imprévisible, ce qui est inférieur aux exigences de la Convention européenne des droits de l'homme.

Enfin, si une action internationale coordonnée peut déboucher sur des mesures efficaces et respectueuses des droits de l'homme, une telle coordination doit mettre en œuvre les droits de l'homme dès la conception et doit impliquer pleinement tous les groupes de parties prenantes tant dans la phase de conception que dans celle de mise en œuvre.

3. La ligne de démarcation entre public et privé

La Convention européenne des droits de l'homme et la jurisprudence de la Cour européenne des droits de l'homme constituent un cadre extraordinairement riche et vivant dans lequel les droits de l'homme peuvent être nourris et protégés. La Convention est contraignante pour les États, sous la forme d'obligations positives et négatives pour la protection de ces droits. Les droits énoncés dans la Convention constituent la norme par défaut, les restrictions étant des exceptions qui doivent être justifiées.

Les gouvernements ont donc deux tâches concernant la modération du contenu en ligne :

- de remplir leurs obligations positives et négatives au titre de la Convention et
- pour assurer l'application du droit national.

Bien respectées, ces deux tâches génèrent une grande synergie entre les obligations de respecter la Convention et de veiller à l'application du droit national. Si les États s'appuient sur des entreprises privées et s'engagent dans des systèmes de corégulation, ils doivent veiller à ce que les principes ou les conditions des restrictions soient respectés (par exemple en ce qui concerne la prévisibilité et la proportionnalité).[68] Il a été avancé que les garanties en matière de droits de l'homme améliorent la qualité du droit. [69]

Un projet de corégulation qui cherche à garantir la prévisibilité et la proportionnalité aura besoin d'outils pour évaluer l'efficacité et la proportionnalité initiale et continue de la mesure. Ceci, à son tour, conduira à une application plus efficace et plus ciblée de la loi et permettra aux décideurs politiques de suivre l'évolution du problème traité. Un projet d'autorégulation ou de corégulation qui ne comporte pas d'objectifs, d'options d'adaptation ou d'abandon, ou de mécanismes de contrôle indépendants ne pourra pas faire la preuve de sa proportionnalité ni de son efficacité.

Dans son rapport du 16 mai 2011, le rapporteur spécial des Nations unies sur la liberté d'opinion et d'expression a déclaré que "les mesures de censure ne devraient jamais être déléguées à une entité privée".[70] Il a donné l'exemple d'une "entité quasi étatique et quasi privée chargée de réglementer les

[68] Voir *Costello-Roberts c. Royaume-Uni*, no. 13134/87, 25 mars 1993, para. 27 : "l'État ne peut pas s'exonérer de sa responsabilité en déléguant ses obligations à des organismes ou des personnes privées", cité dans Kuczerawy A. (2017) "The power of positive thinking : intermediary liability and the effective enjoyment of the right to freedom of expression", https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3033799 , consulté le 5 janvier 2021.

[69] Husovec M. (2021), op cit., p.12.

[70] Conseil des droits de l'homme des Nations unies, Rapport du rapporteur spécial sur la promotion et la protection du droit à la liberté d'opinion et d'expression, Frank La Rue, 16 mai 2011, www.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf , consulté le 6 mai 2020. Voir également

contenus en ligne" comme exemple de ce type de délégation de pouvoir. Une telle délégation se fait parfois par le biais de pressions directes ou indirectes des Etats sur les intermédiaires d'internet. Néanmoins, de nombreux intermédiaires d'internet peuvent restreindre le contenu dans le cadre de leur liberté contractuelle et le font légitimement. Une restriction volontaire, telle que la limitation ou la suppression de types de contenus clairement définis sur une plateforme en ligne destinée aux enfants ou à une profession particulière (une plateforme de discussion de questions médicales, par exemple) ne poserait normalement pas de problèmes pour la liberté d'expression.

La ligne de démarcation entre la corégulation légitime, la pression illégitime de l'État sur les intermédiaires d'internet et l'application légitime par les intermédiaires d'internet de leurs règles internes a été notoirement difficile à tracer. Le risque de censure privatisée est particulièrement grave dans les situations où les États ont mis en place un régime de responsabilité qui incite les intermédiaires d'internet à imposer des restrictions contractuelles nominales, ce qui peut être encore aggravé si la loi mise en œuvre n'est pas claire en soi. Les États, individuellement et collectivement, ont l'obligation positive d'atténuer ce risque en assurant un cadre juridique clair et prévisible, dans lequel les lois interdisant certains types de contenu et les règles de responsabilité sont claires.

Il existe un champ considérable de restrictions mal définies et potentiellement contre-productives (tant du point de vue des droits de l'homme que des objectifs de politique publique visés) imposées par les intermédiaires d'internet, sous la pression directe (par exemple en exigeant un code de conduite) ou indirecte (par exemple en appliquant ou même en ayant des règles de responsabilité excessives) des gouvernements. Le rapporteur spécial des Nations unies a donné un exemple clair de la manière dont les actions des États peuvent conduire à des restrictions des droits de l'homme qui équivalent à une "censure par procuration"[71] :

> *"Toutefois, si un système de notification et de retrait est un moyen d'empêcher les intermédiaires de se livrer activement à des comportements illicites ou de les encourager sur leurs services, il est sujet à des abus de la part des acteurs tant publics que privés. Les utilisateurs qui sont informés par le fournisseur de services que leur contenu a été signalé comme illégal ont souvent peu de recours ou peu de ressources pour contester le retrait. En outre, étant donné que les intermédiaires peuvent toujours être tenus pour financièrement ou dans certains cas pénalement responsables s'ils ne retirent pas le contenu à la réception de la notification des utilisateurs concernant le contenu illégal, ils sont incités à pécher par excès de sécurité en sur-censurant le contenu potentiellement illégal. Le manque de transparence du processus décisionnel des intermédiaires masque souvent aussi les pratiques discriminatoires ou les pressions politiques qui affectent les décisions des entreprises. En outre, les intermédiaires, en tant qu'entités privées, ne sont pas les mieux placés pour déterminer si un contenu particulier est illégal, ce qui nécessite un équilibre minutieux entre les intérêts concurrents et l'examen des moyens de défense*. " [72]

Conseil des droits de l'homme des Nations unies, Rapport du rapporteur spécial sur la promotion et la protection du droit à la liberté d'opinion et d'expression, David Kaye, 6 avril 2018, www.undocs.org/A/HRC/38/35 , consulté le 4 janvier 2020.

[71] Kuczerawy Aleksandra (2018) "Private enforcement of public policy : freedom of expression in the era of online gatekeeping", Liras.

[72] Conseil des droits de l'homme des Nations unies, Rapport du rapporteur spécial sur la promotion et la protection du droit à la liberté d'opinion et d'expression, Frank La Rue, 16 mai 2011, para. 42, www.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf , consulté le 6 mai 2020.

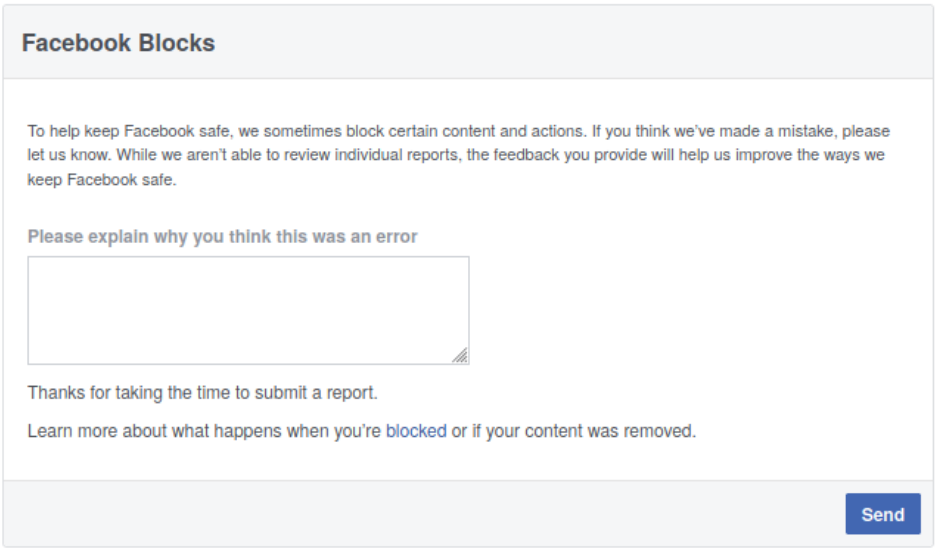


Figure 1: Message reçu par un utilisateur de Facebook en réponse à une plainte selon laquelle le site empêche les utilisateurs de publier des liens vers son site web entièrement légal.

L'ampleur des restrictions non fondées sur le droit imposées par des sociétés privées pour des raisons ostensiblement d'ordre public a également été mise en évidence dans une étude réalisée par l'Institut suisse de droit comparé pour le Conseil de l'Europe.[73] Cette recherche a identifié un blocage généralisé des ressources en ligne par les fournisseurs d'accès à Internet sans base juridique claire dans le droit national.[74] Ce blocage s'est fait sur la base d'initiatives dites d'autorégulation et concernait souvent des ressources en ligne dont le contenu lui-même n'était pas illégal (prétendues violations du droit d'auteur) ou qui constituaient des preuves de comportement criminel (matériel pédopornographique). Dans ce dernier cas, le blocage était effectué sans traiter ces informations avec le sérieux qu'elles méritaient, compte tenu du fait qu'elles étaient censées constituer la preuve d'un délit grave.

Plus il est urgent de s'engager dans de telles restrictions, par exemple en ce qui concerne le matériel pédopornographique, plus il est important de garantir la responsabilité. La mise en œuvre de mécanismes efficaces de transparence, de révision et d'ajustement est cruciale pour garantir l'efficacité et la proportionnalité. Malheureusement, ce n'est pas toujours le cas. Dans une résolution adoptée à une large majorité (597 voix pour, 6 contre)[75] en 2017, le Parlement européen a été très clair sur la lutte contre le matériel pédopornographique en ligne :

> *"considérant que le rapport de mise en œuvre de la Commission ne fournit aucune statistique sur le retrait et le blocage des sites web contenant ou diffusant des images d'abus pédosexuels, en particulier des statistiques sur la rapidité du retrait des contenus, la fréquence du suivi des rapports par les autorités répressives, les retards dans les retraits dus à la nécessité d'éviter*

[73] Conseil de l'Europe (2017) "Étude comparative sur le blocage, le filtrage et le retrait des contenus Internet illégaux", https://edoc.coe.int/en/internet/7289-pdf-comparative-study-on-blocking-filtering-and-take-down-of-illegal-internet-content-.html , consulté le 4 mai 2020.

[74] Il ne s'agit pas, à proprement parler, de "modération de contenu" au sens habituel du terme. Néanmoins, elle mérite d'être soulignée dans ce contexte en raison des nombreuses caractéristiques qui se recoupent avec la modération du contenu.

[75] Voir https://oeil.secure.europarl.europa.eu/oeil/popups/sda.do?id=30474&l=en , consulté le 12 janvier 2021.

toute interférence avec les enquêtes en cours, ou la fréquence à laquelle les données stockées sont effectivement utilisées par les autorités judiciaires ou répressives ;". [76]

a) Imprévisibilité des restrictions de contenu

La mise en œuvre de restrictions sur la base des règles internes d'un intermédiaire d'internet limite la liberté d'expression des individus. La Convention européenne des droits de l'homme exige que les restrictions à ses articles 8 à 11 soient prévues par la loi. Cela ne signifie pas que les restrictions ne peuvent pas être imposées par les sociétés Internet, mais cela signifie que ces règles doivent être claires et appliquées de manière cohérente afin que les utilisateurs puissent adapter leur comportement.[77] Cela est particulièrement vrai pour les restrictions mises en œuvre avec la coopération de l'État dans le cadre de procédures de corégulation, où les niveaux de clarté devraient respecter les exigences de prévisibilité prévues par la Convention et la jurisprudence pertinente et ne devraient pas équivaloir à une délégation abusive de pouvoirs de censure à des parties privées.

Comme nous l'avons vu plus haut, dans le cas des approches d'autorégulation et de corégulation, les documents contractuels (appelés selon les cas "conditions de service", "directives communautaires", etc.) définissent les limites de ce qui est ou n'est pas autorisé et sont généralement appliqués dans le cadre des restrictions d'autorégulation et de corégulation.

Une étude réalisée en 2017 pour le Conseil de l'Europe par le FGV Direito Rio s'est penchée sur la question des droits de l'homme et des contrats de plate-forme en ligne.[78] Elle décrit les conditions de service comme suit :

> *"Les conditions de service sont des contrats standardisés, définis unilatéralement et offerts sans discrimination et à des conditions égales à tout utilisateur. Comme les utilisateurs n'ont pas le choix de négocier, mais seulement d'accepter ou de refuser ces conditions, les conditions de service font partie de la catégorie juridique des accords d'adhésion. En fait, ces accords établissent une sorte de relation "à prendre ou à laisser", qui remplace le concept traditionnel de clauses négociées entre les parties contractantes*. ” [79]

Les recherches menées dans le cadre de cette étude ont révélé que "ces termes sont généralement longs, denses et formulés dans un langage difficile à comprendre pour quiconque n'a pas de formation juridique [...] les gens ne lisent presque jamais ces contrats [...]. Lorsqu'ils le font, ils les trouvent difficiles à comprendre. ” [80]

---

[76] Résolution du Parlement européen du 14 décembre 2017 sur la mise en œuvre de la directive 2011/93/UE du Parlement européen et du Conseil du 13 décembre 2011 relative à la lutte contre les abus sexuels et l'exploitation sexuelle des enfants et la pédopornographie, www.europarl.europa.eu/doceo/document/TA-8-2017-0501_EN.html , consulté le 27 mai 2020.

[77] *Centro Europa 7 S.R.L. et Di Stefano c. Italie* (GC), no. 38433/09, 7 juin 2012, para. 141 : « Ainsi, une norme ne peut être considérée comme une "loi" que si elle est formulée avec suffisamment de précision pour permettre aux citoyens de régler leur conduite ; ils doivent pouvoir - le cas échéant avec des conseils appropriés - prévoir, dans une mesure raisonnable dans les circonstances, les conséquences qu'une action donnée peut entraîner ».

[78] Venturini J., Louzada L., Maciel M.F,, Zingales N., Stylianou K., Belli L. (2016) *Terms of service and human rights : an analysis of online platform contracts*, Editora Revan, ISBN 978-85-7106-574-1, https://internet-governance.fgv.br/sites/internet-governance.fgv.br/files/publicacoes/terms_of_services_06_12_2016.pdf , consulté le 8 juin 2020.

[79] Venturini et al (2016), op cit., p.23, citant Lemley M. A. (2006) *Conditions d'utilisation*, 91 MINN. L. REV. 459, 459, www.kentlaw.edu/faculty/rwarner/classes/ecommerce/2008/contracts/consent/lemley%20tersm%20of%20use.pdf , consulté le 6 mai 2021.

[80] Ibid, p.24.

À titre d'exemple de ce que cela signifie pour la prévisibilité des restrictions de contenu, l'étude a révélé que, sur 50 fournisseurs de services, 23 avaient des dispositions contradictoires (20) ou aucune (3) disposition dans leurs conditions de service quant à savoir s'ils analysent, bloquent ou suppriment des contenus pour "des raisons quelque peu spécifiques, indéterminées ou peu claires".[81]

Il est donc clair que la prévisibilité de la signification et de l'application des règles internes des sociétés Internet est souvent inférieure aux normes nécessaires pour garantir la protection des droits de l'homme. Les défaillances doivent être identifiées et corrigées avant le lancement de tout système d'autorégulation ou de corégulation reposant sur ces règles.

b) Manque de clarté sur l'équilibre des rôles et des responsabilités des États et des acteurs privés

Ces considérations créent un ensemble très complexe de critères permettant d'évaluer le respect des obligations positives et négatives des États en matière de restrictions des droits de l'homme par des mesures d'autorégulation et de corégulation.

Comme l'a souligné le rapporteur spécial des Nations unies, même un système aussi superficiellement non controversé que celui de "notification et retrait", tel qu'il est largement utilisé dans la région du Conseil de l'Europe, peut ne pas offrir une protection suffisante de la liberté d'expression.[82]

Les restrictions sont parfois mises en œuvre en raison du manque de clarté des lois sur la responsabilité et parfois comme moyen de contourner le problème que les lois sur l'illégalité des contenus sont elles-mêmes peu claires. Par exemple, un projet financé par la Commission européenne a constaté qu'il existait "d'énormes disparités" dans l'UE entre les lois nationales interdisant le racisme et la xénophobie, bien qu'elles soient basées sur la législation européenne.[83]

Face à la disparité des lois nationales en Europe, même l'agence de police européenne Europol est légalement tenue de signaler les contenus en ligne potentiellement liés à des crimes graves comme d'éventuelles violations des conditions de service "pour la considération volontaire" des intermédiaires d'internet, plutôt qu'une éventuelle violation de la loi.[84] Le contenu potentiellement illégal en question est apparemment suffisamment grave pour justifier une action de la part d'Europol, mais pas assez pour que les entreprises soient tenues de prendre des mesures ou soient définitivement mises hors la loi en vertu du droit européen.

La situation très floue dans laquelle se trouvent les intermédiaires d'internet a été mise à nu par une réponse à une question parlementaire sur ce sujet posée à la Commission européenne.[85] Comme la perte de l'immunité de responsabilité est déclenchée lorsque l'intermédiaire d'internet acquiert une "connaissance effective" d'un contenu illégal, un parlementaire a demandé si une référence d'Europol constituerait une telle connaissance. La Commission européenne a répondu que cela ne constituerait

[81] Ibid, p.54.

[82] Le rapporteur spécial des Nations unies (ONU) sur la liberté d'opinion et d'expression, le représentant de l'Organisation pour la sécurité et la coopération en Europe (OSCE) sur la liberté des médias, le rapporteur spécial de l'Organisation des États américains (OEA) sur la liberté d'expression et le rapporteur spécial de la Commission africaine des droits de l'homme et des peuples (CADHP) sur la liberté d'expression et l'accès à l'information, "Déclaration conjointe sur la liberté d'expression et l'Internet", juin 2011, para. 2b, www.oas.org/en/iachr/expression/showarticle.asp?artID=848 , consulté le 4 janvier 2020.

[83] Projet Mandola, "Définition de la haine illégale et de ses implications", 31 mars 2016, page 7, http://mandola-project.eu/m/filer_public/7b/8f/7b8f3f88-2270-47ed-8791-8fbfb320b755/mandola-d21.pdf , consulté le 15 mai 2020.

[84] Article 4(1)m du règlement 2016/794 du Parlement européen et du Conseil, 11 mai 2016, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0794 , consulté le 7 mai 2020.

[85] Question E-7205/17 de Cornelia Ernst à la Commission européenne, 23 novembre 2017, www.europarl.europa.eu/doceo/document/E-8-2017-007205_EN.html , consulté le 5 janvier 2021.

pas une "connaissance effective" à moins que le fournisseur n'ait connaissance "de faits et de circonstances sur la base desquels un opérateur économique diligent aurait dû identifier l'illégalité en question". "En d'autres termes, même un renvoi d'Europol ne suffirait pas nécessairement à déclencher la connaissance de l'illégalité, même s'il est possible que, dans certaines circonstances indéfinies[86], ce soit le cas. On ne voit[87] pas bien pourquoi, dans de telles circonstances, un service chargé d'assurer l'application de la loi ne serait pas tenu de prendre les mesures appropriées en coopération avec les autorités judiciaires et les organes répressifs nationaux plutôt que d'avoir l'obligation de se fier aux conditions de service d'un intermédiaire d'internet.

Dans une situation où l'intermédiaire d'internet reçoit un renvoi d'Europol, plusieurs facteurs l'incitent à retirer le contenu :

- le fait que le renvoi provenait d'un organisme chargé de l'application de la loi (bien que sans notification explicite que le contenu était illégal) ;
- les éventuels préjudices de réputation liés au fait de ne pas répondre à ces renvois en supprimant le contenu en question ;
- le fait que le renvoi peut (ou non) être interprété par un tribunal comme une "connaissance effective" ou une "connaissance" de l'illégalité, créant ainsi une responsabilité pour l'intermédiaire d'internet, même si Europol n'a pas pu ou n'a pas voulu fournir - par le biais d'un renvoi aux autorités judiciaires nationales, par exemple - la confirmation de l'illégalité du contenu.

Toutefois, il n'y a pas grand-chose d'autre que la relation commerciale de l'entreprise avec un client individuel, qui l'incite à garder le contenu en ligne.

Dans un tel environnement, qui est en faute si les opérateurs retirent régulièrement, comme ils y sont incités, des contenus légaux, sans que cela soit clairement prévisible, nécessaire ou proportionné ?

Il y a deux possibilités. Dans le premier cas de figure, la décision est privée, fondée sur des règles privées que les individus ont acceptées lorsqu'ils ont eu la possibilité de « prendre ou de laisser ». Dans ce cas, l'action sort éventuellement du champ d'application du droit des droits de l'homme et relève de la responsabilité combinée de l'intermédiaire d'internet et de l'utilisateur. Cet argument apparait comme non recevable à la lumière de la jurisprudence de la Cour européenne des Droits de l'Homme. Par exemple, dans l'affaire *Cengiz et autres c. Turquie*, l'internet a été décrit comme "l'un des principaux moyens par lesquels les individus exercent leur droit à la liberté de recevoir et de communiquer des informations et des idées, car il fournit des outils essentiels pour participer à des activités et à des discussions concernant des questions politiques et des questions d'intérêt général".[88]

Dans le deuxième cas, ce sont les États qui

- n'ont pas réussi, individuellement et collectivement, à exiger que les conditions de service soient claires ;
- n'ont pas exigé que les conditions de service soient mises en œuvre de manière prévisible et proportionnée ;

[86] La formulation de la réponse de la Commission européenne provient d'une affaire de la Cour de justice des Communautés européennes relative à la vente non autorisée de cosmétiques et non à un contenu illicite, affaire 324/09, L'Oréal SA et autres contre eBay International AG et autres, http://curia.europa.eu/juris/document/document.jsf?text=&docid=107261&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=11354812 , consulté le 27 juillet 2020.

[87] Question parlementaire E-7205/2017, www.europarl.europa.eu/doceo/document/E-8-2017-007205_EN.html , consulté le 20 mai 2020.

[88] *Cengiz et autres c. Turquie*, nos. 48226/10 et 14027/11, 1 decembre 2015, para. 49.

- ont omis d'adopter des lois claires au niveau national ou international concernant le contenu illégal en question ;
- n'ont pas veillé à ce qu'une autorité compétente soit en place pour émettre un ordre de retrait ou de non-retrait du contenu en question et
- n'ont pas fourni un cadre juridique garantissant un équilibre plus approprié des incitations pour les prestataires.

Si c'est le cas, cela ne respecte pas les obligations positives des États en vertu de la Convention européenne des droits de l'homme et est contraire à l'exigence selon laquelle toute restriction doit être prévue par la loi.

Il est clair, cependant, que les grands intermédiaires d'internet sont mieux équipés que leurs petits concurrents pour faire face à cette incertitude juridique. De même, les grands intermédiaires d'internet sont mieux placés que leurs petits concurrents pour faire face à des règles de responsabilité plus onéreuses (par exemple en achetant ou en développant des technologies de filtrage pour respecter des délais très courts pour le retrait des contenus). À l'heure où les décideurs politiques expriment de plus en plus de préoccupations quant à la puissance croissante des plus grandes plateformes, cela devrait être une considération importante pour les États. [89]Par conséquent, dans l'intérêt des droits de l'homme et aussi pour garantir l'innovation et la concurrence, il convient absolument d'éviter les règles de responsabilité peu claires ou onéreuses.

En conclusion, il incombe à l'État de veiller à ce que les règles de responsabilité des intermédiaires d'internet soient suffisamment claires pour éviter d'encourager une censure privatisée et à ce que les règles internes des intermédiaires d'internet, ainsi que leur mise en œuvre, soient claires, prévisibles et proportionnées. Dans le cas contraire, les États incitent à des restrictions de la liberté d'expression, ne respectant pas les garanties prévues par la Convention au sujet des restrictions autorisées *et* permettent que cela se produise dans un environnement où les individus sous leur protection sont sans défense contre les caprices des intermédiaires d'internet à but lucratif.

c) Les risques de surconformité, en particulier lorsqu'elle conduit à des résultats discriminatoires

La surconformité :

Dans un environnement où un intermédiaire d'internet peut être tenu responsable lorsqu'il ne retire pas un contenu ou des services qui pourraient constituer une infraction, mais où il a peu d'incitations à maintenir le contenu en ligne, il semble inévitable que les intermédiaires d'internet restreignent le contenu qui tombe dans la "zone grise", en se fondant sur leurs conditions de service pour ce faire.[90] En conséquence, les idées, explicitement protégées par la Convention européenne des droits de l'homme, comme celles "qui offensent, choquent ou perturbent l'État ou une partie de la population", [91]ont peu de chances d'être protégées dans la pratique. Les recherches indiquent que "les fournisseurs ont tendance à retirer trop de contenu pour éviter toute responsabilité et économiser des ressources,

[89] Sweney Mark, "Google and Facebook dominance should be curted, suggests CMA", *The Guardian*, 18 décembre 2019, www.theguardian.com/business/2019/dec/18/google-facebook-dominance-curbed-cma-report-uk-digital-market , consulté le 9 juillet 2020.

[90] Le manque de transparence signifie qu'il y a peu de données empiriques à ce sujet. Cependant, les exemples sont si nombreux que ce phénomène semble difficile à nier. Voir, par exemple, "YouTube efface l'histoire", *New York Times*, 23 octobre 2019, www.nytimes.com/2019/10/23/opinion/syria-youtube-content-moderation.html , consulté le 18 janvier 2020.

[91] *Handyside c. Royaume-Uni*, no. 5493/72, 7 décembre 1976, para. 49.

ils utilisent également la technologie pour évaluer les notifications ; et les utilisateurs concernés qui ont publié le contenu n'entreprennent souvent aucune action".[92]

Discrimination :

En outre, les questions de discrimination et des tentatives à « déjouer » les systèmes (c'est-à-dire la manipulation délibérée, pas nécessairement illégale, ou l'abus des systèmes de plainte des intermédiaires d'internet) se posent dans un système qui repose largement sur des approches d'autorégulation et de corégulation pour traiter les contenus en ligne indésirables ou illégaux et qui ne dispose pas d'un cadre juridique clair pour ces restrictions.

En l'absence d'une obligation de le faire, il n'y a aucune raison pour qu'un opérateur, accidentellement ou délibérément, ne donne pas la priorité au traitement des abus racistes contre un groupe plutôt qu'un autre, ou d'une forme de sexisme plutôt qu'une autre.[93] En effet, en l'absence de règles claires en matière de transparence, notamment en cas d'utilisation de l'intelligence artificielle, il serait impossible de savoir si l'approche discriminatoire a même eu lieu. Il semble contre-intuitif de s'appuyer de plus en plus sur l'intelligence artificielle pour prendre des décisions parfois très complexes sur la légalité, le respect des conditions de service et le contexte de la parole. Pourtant, ces technologies sophistiquées produisent rarement, voire jamais, des rapports complets et opportuns sur les raisons spécifiques pour lesquelles le contenu a été retiré (ou non), ni des détails granulaires sur la quantité de contenu qui a été restreinte pour ces raisons.

> *"Il peut être difficile, voire impossible, de déterminer si un système d'IA a un impact sur les droits de l'homme, la démocratie et l'État de droit lorsqu'il n'y a pas de transparence sur l'utilisation d'un système d'IA par un produit ou un service et, si oui, sur la base de quels critères il fonctionne. En outre, sans cette information, une décision prise par un système d'IA ne peut être efficacement contestée, ni le système amélioré ou corrigé lorsqu'il cause un préjudice*. " [94]

Pourtant, il existe de nombreuses preuves tentatives à « déjouer » les systèmes de plaintes des intermédiaires d'internet d'une façon discriminatoire.[95] Sans la transparence nécessaire pour identifier les erreurs et faciliter l'analyse des raisons pour lesquelles des erreurs se sont produites, les problèmes semblent destinés à perdurer. Un exemple particulièrement flagrant est le cas du groupe de défense des droits des femmes "Women on Waves". YouTube a supprimé la chaîne entièrement légale et conforme aux conditions de service de l'organisation à quatre reprises, sans explication, au cours de l'année 2018.[96] Un autre exemple est le retrait temporaire de la page antiraciste "Kick Out

---

[92] Alexandre De Streel et al (2020) "Online Platforms' Moderation of Illegal Online Content", Étude réalisée pour la commission du marché intérieur et de la protection des consommateurs du Parlement européen, www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718_FR.pdf , consulté le 18 janvier 2021.

[93] Maarten Sap et al, "The Risk of Racial Bias in Hate Speech Detection," Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1668-1678, Florence, Italie, 28 juillet - 2 août 2019.

[94] Comité ad hoc sur l'intelligence artificielle du Conseil de l'Europe (2020) "Étude de faisabilité", https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da , consulté le 21 janvier 2021.

[95] Les contenus provenant de ou concernant des femmes semblent en souffrir de manière disproportionnée, allant du fabricant de soutiens-gorge Thirdlove, interdit sur Facebook pour avoir affiché une boîte de soutiens-gorge (https://kernelmag.dailydot.com/issue-sections/features-issue-sections/12796/facebook-nudity-breasts-advertising/, consulté le 14 février 2020), à une publicité sur le blocage des menstruations en raison d'un "excès de peau" (www.independent.co.uk/news/media/online/woman-gets-censored-by-facebook-because-she-blogs-about-periods-a6725176.html, consulté le 14 février 2020). Voir également la note de bas de page suivante.

[96] Austin Evelyn, "Les trois suspensions de YouTube de Women on Waves cette année montrent une fois de plus que nous ne pouvons pas laisser les sociétés Internet contrôler notre discours", Bits of Freedom, juin 2018, www.bitsoffreedom.nl/2018/06/28/women-on-waves-three-youtube-suspensions-this-year-show-yet-again-that-we-cant-let-internet-companies-police-our-speech/ , consulté le 14 février 2020.

Zwarte Piet" d'Instagram alors qu'une page qui aurait incité à la violence contre ce mouvement est restée en ligne.[97]

Les intermédiaires d'internet peuvent, non sans justification, faire valoir que le manque de transparence sur la manière dont ces décisions sont prises est nécessaire pour empêcher les tentatives abusives à « déjouer » les systèmes de plaintes. Toutefois, l'utilisation du manque de transparence comme moyen manifeste de protéger l'intégrité de procédures internes défaillantes externalise le coût de ces procédures (non prouvées) pour les victimes de décisions incorrectes et incohérentes, qui n'ont aucun moyen d'éviter des décisions similaires à l'avenir et ne disposent d'aucun recours.

Les États devraient prendre des mesures, le cas échéant, contre la surconformité et la discrimination en mettant en œuvre des règles de transparence et en fournissant des orientations ou des règles claires sur la prévisibilité et l'équilibre des conditions des contrats de service entre les intermédiaires d'internet et leurs utilisateurs ainsi que sur l'exécution de ces contrats.

d) Se concentrer sur des mesures limitées à la vitesse et au volume, en fonction de situations "urgentes" répétées

La technologie, la criminalité et la société sont en constante évolution. Par conséquent, il est inévitable que l'environnement dans lequel une mesure d'autorégulation ou de corégulation est mise en œuvre change à court ou moyen terme. Il est donc crucial de veiller à ce que les objectifs soient atteints et que des ajustements puissent être apportés à toute mesure, en fonction de l'évolution de la situation.

Trop souvent, lorsqu'un sujet particulier fait la une des journaux, les États et les acteurs privés doivent être perçus comme agissant de manière décisive. Dans ces circonstances, il est facile pour les États d'exiger des intermédiaires d'internet qu'ils fassent "plus" pour lutter contre le problème. Les mesures, comme le simple nombre de messages supprimés, par exemple, incitent à supprimer "plus" et "plus vite". Il est facile de supprimer rapidement un contenu. Cependant, cela entraîne des coûts imprévisibles mais incontestables.

Le premier rapport de mise en œuvre du code de conduite de l'UE sur la lutte contre les discours de haine illégaux en ligne fournit un modèle utile pour démontrer certains défis, dont la résolution est laissée aux intermédiaires d'internet.

Les tâches que le code de conduite était censé accomplir étaient les suivantes :

- de s'attaquer au discours de haine tel que défini par la décision-cadre de 2008 sur la lutte contre certaines formes et expressions de racisme et de xénophobie au moyen du droit pénal[98] et d'empêcher sa propagation, malgré les "énormes disparités" entre les lois mettant en œuvre cette législation[99] et
- pour défendre la liberté d'expression.

D'un point de vue positif, le code peut être considéré comme un effort réussi pour déplacer le débat en dehors des discussions politiques et à des niveaux plus élevés des structures de gestion des intermédiaires d'internet.

---

[97] Hulsen Stan, "Instagram-account van KOZP tijdelijk geschorst, haataccount nogwel online" (compte Instagram du KOZP temporairement suspendu, le compte haineux reste en ligne) Nu.nl, 20 novembre 2019, www.nu.nl/tech/6012319/instagram-account-van-kozp-tijdelijk-geschorst-haataccount-nog-wel-online.html , consulté le 22 mai 2020 (en néerlandais).

[98] Décision-cadre 2008/913/JAI du Conseil, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM%3Al33178, consulté le 8 juin 2020.

[99] Projet Mandola (2016), op cit., p. 9.

Toutefois, en ce qui concerne les mécanismes de supervision et d'ajustement, qui n'étaient peut-être pas au premier plan des réflexions lorsque la mesure a été mise en place, elle fournit une étude de cas utile concernant les pièges des approches d'autorégulation et de corégulation sous la pression des gouvernements.[100]

Six mois après l'adoption du code, un rapport de mise en œuvre intitulé "Code de conduite pour la lutte contre les discours de haine illégaux sur Internet" a été publié : Premiers résultats de la mise en œuvre" a été publié par la Commission européenne.[101]

Sous une rubrique ("méthodologie"), elle fait référence aux notifications évaluées comme "notification de discours de haine *présumés* illégaux" (c'est nous qui soulignons). Sous une autre rubrique ("notifications de discours haineux illégaux faits à des sociétés informatiques"), les mêmes rapports ont été qualifiés de "notifications de discours haineux illégaux" (c'est-à-dire que les "allégations" sont apparemment toutes supposées être valables) et cela a été répété sous les rubriques suivantes. Les seules données fournies dans le cadre de l'exercice de contrôle concernent les types de contenu prétendument illégal, le nombre de notifications, la rapidité de traitement des notifications, les types de notifiants et les pourcentages de notifications qui ont conduit à des suppressions pour chaque fournisseur - et non la quantité réelle de contenu illégal correctement identifié qui a été supprimée (par exemple, par un examen indépendant d'échantillons de contenu supprimés et laissés en ligne, avec des points de référence convenus pour les taux d'erreur acceptables).

Si une proportion significative du contenu retiré était effectivement de nature criminelle et si le contenu qui n'a pas été retiré était effectivement légal, alors ces objectifs seraient, au moins en partie, atteints. Cependant, aucune donnée n'est générée en relation avec ces questions. Nous n'apprenons que la vitesse et la quantité des retraits. Même une mesure minimale comme l'échantillonnage aléatoire des décisions de retrait et de non-retrait, qui donnerait une indication de l'impact actuel, permettrait de procéder à des adaptations.

L'impact des mesures sur les crimes commis n'a jamais été évalué. On pourrait soutenir de manière crédible qu'une suppression efficace de contenus de nature criminelle pourrait dissuader certaines personnes de télécharger des discours haineux criminels, ce qui signifie que la mesure pourrait fonctionner. Toutefois, on pourrait également soutenir de manière crédible que le fait de s'appuyer sur le code de conduite crée un certain degré d'impunité, car la pire sanction possible est une éventuelle suppression de contenus qui, autrement, auraient pu donner lieu à des poursuites pénales, en raison de leur illégalité. Les données brutes sur la vitesse et le volume des suppressions de contenus passent à côté de ces nuances et ne permettent pas de suivre l'évolution du problème dans le temps.

En fait, selon le cinquième rapport de suivi du code de conduite, sur les 3 099 éléments de contenu potentiellement illégal qui ont été retirés par les entreprises participantes, 85% n'ont été renvoyés à aucune autorité chargée de l'application de la loi par aucune organisation participante et aucune donnée n'est fournie quant à savoir si des mesures répressives ont été prises concernant les 15% restants.[102]

---

[100] Les initiatives gouvernementales telles que la loi allemande sur l'application des réseaux suggèrent que les États sont de plus en plus conscients des limites de cette approche.

[101] Commission européenne (2016) "Code de conduite pour la lutte contre les discours de haine illégaux sur Internet : Premiers résultats de la mise en œuvre ", https://ec.europa.eu/newsroom/document.cfm?doc_id=40573 , consulté le 8 juin 2020.

[102] Commission européenne (2020), fiche d'information « 5eme évaluation de la mise en œuvre du Code de conduite », https://ec.europa.eu/info/sites/info/files/codeofconduct_2020_factsheet_12.pdf , consulté le 24 juillet 2020.

e) Modérer le risque

Une autre série de questions se pose en ce qui concerne la modération du contenu juridique pour protéger les groupes vulnérables, tels que les enfants. Il est important de souligner que "le risque n'est pas un mal"[103] et que le fait d'éviter le risque peut être lui-même nuisible. Par conséquent, la pression exercée sur les i intermédiaires d'internet pour qu'ils "s'autorégulent" afin de minimiser les risques pour les groupes vulnérables comporte son propre ensemble de dangers. Par exemple, l'Inspection scolaire britannique (OFSTED) a constaté que "les élèves étaient globalement plus vulnérables lorsque les écoles utilisaient des systèmes fermés parce qu'ils n'avaient pas suffisamment d'occasions d'apprendre à évaluer et à gérer les risques par eux-mêmes."[104] Cela ne signifie pas qu'il ne faut pas modérer le contenu pour protéger les groupes vulnérables, mais plutôt qu'il faut des mécanismes de contrôle et de révision très minutieux. La création d'un risque de responsabilité de la part de l'intermédiaire d'internet, qui doit être atténué par l'élimination ou la minimisation des risques perçus par les groupes vulnérables (même si les avantages potentiels peuvent, ou sont susceptibles de, dépasser les dommages potentiels), n'est peut-être pas la manière la plus efficace, proportionnée et ciblée de parvenir à un résultat équilibré.

Par conséquent, dans de telles circonstances, il est crucial, pour éviter des conséquences involontaires, que toute intervention politique ayant pour but de minimiser les risques soit clairement reconnue comme telle, 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 protection significative de la liberté d'expression, ainsi que d'outils permettant d'identifier les effets contre-productifs.

III Droits affectés

Les cadres d'autorégulation et de corégulation pour la modération de contenu prévoient diverses mesures pour rendre l'activité illégale plus difficile. Par exemple, ils peuvent refuser certains mots clés (par exemple "ivoire") d'une publicité en ligne dans une juridiction où la vente du produit est illégale. Les mesures peuvent également porter sur le retrait de contenus, d'activités ou de comptes qui entrent dans le champ d'application du cadre, sur la base de plaintes ou d'avis reçus de tiers. YouTube (propriété de Google), par exemple, supprime globalement les comptes des personnes qui font l'objet de trois plaintes correctement formulées en matière de droits d'auteur en vertu de la législation américaine et la recherche Google rétrograde globalement les domaines qui font l'objet de "grandes quantités" de plaintes correctement formulées.[105]

Ces restrictions ont une incidence sur la liberté d'expression, le droit à la vie privée, la non-discrimination, la liberté de réunion et le droit à un recours effectif, tandis que le fait de ne pas s'attaquer de manière adéquate aux comportements abusifs ou illégaux peut également porter atteinte à toute une série de droits de l'homme.[106]

---

[103] Livingstone Sonia, Kalmus Veronika,Talves Kairi (2014) *Expériences des filles et des garçons en matière de risques et de sécurité en ligne*, dans : Carter Cynthia, Steiner Linda, McLaughlin Lisa (Ed.) *The Routledge Companion to Media and Gender,* (190-200), Routledge, p. 192.

[104] OFSTED (2014) "The Safe Use of New Technologies", p.4, https://webarchive.nationalarchives.gov.uk/20141105221831/https://www.ofsted.gov.uk/sites/default/files/documents/surveys-and-good-practice/t/The%20safe%20use%20of%20new%20technologies.pdf, consulté le 9 juillet 2020.

[105] Google Public Policy Blog, "Continued progress on fighting piracy", 17 octobre 2014, https://publicpolicy.googleblog.com/2014/10/continued-progress-on-fighting-piracy.html , consulté le 9 juillet 2020.

[106] Parmi de nombreux exemples, Netflix aurait utilisé des demandes de retrait de droits d'auteur pour faire taire les critiques d'un film controversé, par exemple. Voir Cox Katie, "Netflix files copyright claims against tweets criticising movie,

1. Liberté d'expression

En raison de l'implication de l'État, un système de corégulation qui restreint la liberté d'expression doit respecter des critères minimums, comme le stipule l'article 10.2 de la Convention européenne des droits de l'homme.

Les questions à prendre en compte pour évaluer la légalité de ces restrictions sont les suivantes

- Y a-t-il eu une ingérence dans le droit en question et, si oui, a-t-elle été prescrite par la loi ?
- Était-elle réellement destinée à poursuivre un ou plusieurs des objectifs légitimes en question ?
- Compte tenu de toutes les circonstances pertinentes, était-il nécessaire et proportionné dans une société démocratique à ces fins ?[107]

Il convient également d'accorder l'attention nécessaire au fait que les restrictions imposées par la modération de contenu se déroulent dans un environnement déjà difficile pour l'exercice de ce droit, en raison des effets paralysants du profilage en ligne omniprésent.

Si, comme nous l'avons mentionné, les approches d'autorégulation et de corégulation présentent l'avantage de pouvoir être conçues et mises en œuvre rapidement, cette rapidité se fait au détriment de la délibération et des contrôles et contrepoids d'une procédure législative ou judiciaire. Cette rapidité ne doit pas se faire au détriment de l'examen de questions fondamentales comme celles-ci. Lorsque la rapidité est essentielle, des règles claires, prévisibles et responsables devraient être mises en place pour garantir que les restrictions puissent être temporairement appliquées en attendant une évaluation finale.

Le code de conduite de la Commission européenne sur la lutte contre les discours de haine en ligne, qui est clairement basé sur une approche d'autorégulation, aborde cette question de manière explicite, en soulignant la nécessité de défendre la liberté d'expression, en utilisant des formulations tirées de la jurisprudence de la Cour européenne des droits de l'homme. Malheureusement, cependant, aucun effort visible n'a été fait dans les rapports de mise en œuvre pour évaluer dans quelle mesure cette disposition a, en fait, été respectée.

Néanmoins, il est très positif que la Commission européenne ait pris l'initiative de préparer des "rapports d'évaluation" pour contrôler l'effet du code, même si les mesures utilisées sont assez limitées.[108]

Il est essentiel, tant avant qu'après la mise en œuvre d'une mesure de corégulation, que le respect de la législation et des principes relatifs aux droits de l'homme soit examiné.

Étude de cas : Pages supprimées / interdictions de l'ombres Au Royaume-Uni, les forums publics ("pages") de huit organisations indépendantes de la société civile ont été supprimés par Facebook le 4 novembre 2019 (au cours d'une campagne électorale

trailer", Arstechnica.com, 11 mai 2020, https://arstechnica.com/tech-policy/2020/11/netflix-dmca-takedown-requests-hit-negative-tweets-about-cuties/ , consulté le 18 janvier 2021.

[107] Greer Steven, "Les exceptions aux articles 8 à 11 de la Convention européenne des droits de l'homme", Éditions du Conseil de l'Europe, 1997, www.echr.coe.int/LibraryDocs/DG2/HRFILES/DG2-EN-HRFILES-15(1997).pdf, consulté le 28 août 2020.

[108] Les rapports de suivi sont disponibles sur le site https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en , consulté le 5 janvier 2021.

générale). Tous ces groupes ont en commun d'avoir commencé comme des organisations pro-UE, d'être tous des groupes locaux de bénévoles, basés dans des villes individuelles et d'avoir une orientation locale et pro-UE qui ressort clairement de leur nom ("Banbury for Europe", par exemple). Les groupes font aussi généralement campagne hors ligne.

Certains de ces groupes ont également fait l'objet d'"interdictions de l'ombre" répétées (qui laissent le contenu/les comptes des groupes en ligne mais les rendent nettement plus difficiles à trouver) en novembre et décembre de cette année-là. L'impact de ces mesures a été une réduction de la "portée quotidienne" des pages de plus de 90%. L'impact des actions de Facebook les actions et le succès des groupes en en termes d'influence sur l'élection dans les circonscriptions où ils étaient actifs est difficile à savoir.

Aucune accusation d'activité illégale n'a été faite, ni aucune allégation spécifique de violation des « conditions de service » de Facebook. Facebook a suggéré que les groupes modifient leur comportement, par exemple en réduisant le nombre de messages sur les pages. Cependant, Facebook n'a pas dit si cela permettrait, en fait, d'éviter que le même problème ne se reproduise. La société a expliqué que cette restriction de la liberté d'expression des groupes "est automatiquement prise par notre intelligence artificielle suite à l'activité entreprise par la page".

Dans le contexte d'une plate-forme en ligne qui est maintenant si centrale pour les communications en ligne, cela soulève de multiples questions concernant les obligations positives de l'État à garantir la liberté d'expression, et les niveaux minimums de transparence, de prévisibilité, d'équité et de réparation que l'on peut raisonnablement attendre d'une plate-forme Internet.[109] Cela soulève également une autre question : si nous attendons de l'intelligence artificielle qu'elle soit suffisamment performante pour interpréter un langage complexe et son contexte et prendre des décisions basées sur cette analyse, nous devons également exiger qu'elle soit à même de fournir automatiquement des données significatives sur la base de cette interprétation.

2. Droit à la vie privée

La modération du contenu nécessite le traitement d'une série de données personnelles. Par exemple, afin de mettre en œuvre des mesures telles que la politique de "trois fautes" de YouTube, une série de données personnelles et non personnelles doivent être stockées par l'entreprise, telles que le nom d'utilisateur de la personne, le nom du plaignant, la justification du retrait du contenu, les dates et heures des téléchargements et des retraits, etc.

En outre, le traitement de ces données peut inclure le traitement de catégories spéciales de données telles que celles relatives aux opinions politiques, à l'appartenance à un syndicat, aux croyances religieuses ou autres. Ces données ne peuvent être traitées dans le cadre de la Convention 108(+)[110] que si des garanties appropriées existent en droit. Il serait utile que les États membres du Conseil de l'Europe examinent s'il existe des motifs juridiques spécifiques pour le traitement de données à caractère personnel en relation avec tous les aspects de la modération du contenu.

Le droit à la vie privée des plaignants doit faire l'objet d'une attention particulière, notamment en ce qui concerne les signalements de personnes ou de groupes vulnérables. Tout manquement à cette

[109] Horten Monica (2020) "Algorithms Patrolling Content : Où est le mal ? ", https://ssrn.com/abstract=3792097 , consulté le 6 mai 2021.

[110] Protocole d'amendement à la Convention pour la protection des personnes à l'égard du traitement des données à caractère personnel, adopté par le Comité des Ministres lors de sa 128eme session à Elseneur le 18 mai 2018.

règle peut entraîner des répercussions/une réaction à l'encontre de ces personnes et groupes, leur causant un préjudice direct et ayant un effet dissuasif sur les futurs signalements.

Par ailleurs, tout en assurant autant de transparence aux utilisateurs que possible, une attention particulière devrait être accordée à tout texte/image de remplacement qui apparaîtrait à la place du contenu qui a été supprimé. En l'absence d'une décision juridiquement contraignante, il pourrait ne pas être approprié d'utiliser une formulation qui pourrait être comprise comme une accusation d'infraction à la loi, en particulier si la loi appliquée est celle d'un pays autre que le lieu de résidence de l'auteur du téléchargement. En Europe, Google balise toutes les recherches de noms de personnes (mais uniquement lorsque le prénom et le nom de famille sont tous deux utilisés) avec un avis confus et inutile disant "certains résultats peuvent avoir été supprimés en vertu de la législation sur la protection des données en Europe", qui ne fournit aucune information utile à ceux qui le lisent, comme par exemple pourquoi cela pourrait être le cas ou quelle est la probabilité que ce soit le cas. Toute notification doit être claire et significative pour les utilisateurs.

Étude de cas : Placeholders et protection des données

À titre d'exemple des problèmes à éviter, l'expérience des utilisateurs du service d'hébergement web mooo.com mérite d'être soulignée. Mooo.com était un service d'hébergement de sites web, où chaque site hébergé constituait un sous-domaine du nom de domaine du service (hostedsite.mooo.com, par exemple). Un petit nombre de ces sites hébergés ont été découverts comme contenant du matériel illégal.[111]Au lieu de saisir le petit nombre de sites contenant du matériel illégal, les autorités américaines ont saisi la totalité du domaine mooo.com et ont remplacé tout ce qu'il hébergeait par l'image ci-dessous. En conséquence, les visiteurs de l'un des 84 000 sites web légaux hébergés sur le service ont vu l'image, même si le site qu'ils visitaient n'avait jamais contenu de matériel illégal.

Figure 2Image présentée aux visiteurs de dizaines de milliers de sites web entièrement légitimes hébergés sur mooo.com, suite à la saisie du domaine par les autorités policières américaines

[111] Van Der Sar Ernesto, "US Government shuts down 84,000 websites 'by mistake'", torrentfreak.com, 16 février 2011, https://torrentfreak.com/u-s-government-shuts-down-84000-websites-by-mistake-110216/ , consulté le 27 mai 2020.

3. Liberté de réunion et d'association

Les questions relatives à la modération du contenu et à la liberté de réunion sont expliquées dans le commentaire général 37 du Comité des droits de l'homme des Nations unies.[112] Ce document explique au paragraphe 9 que "la protection complète du droit de réunion pacifique n'est possible que lorsque d'autres droits, qui se recoupent souvent, sont également protégés, notamment la liberté d'expression, la liberté d'association et la participation politique". Il souligne également, au paragraphe 34, l'obligation positive des États de veiller à ce que "les fournisseurs de services Internet et les intermédiaires d'internet ne restreignent pas indûment les rassemblements ou la vie privée des participants aux rassemblements", ce qui couvrirait à la fois les mesures d'autorégulation et de corégulation.

La liberté de réunion et d'association peut être compromise de deux manières par une modération inappropriée du contenu, à savoir hors ligne ou en ligne. La planification en ligne de manifestations physiques peut être entravée et, deuxièmement, la liberté de réunion et d'association en ligne peut être restreinte.

Étude de cas - liberté de réunion et d'association :

Quatre grandes organisations de défense du climat (et, semble-t-il, des centaines d'autres) ont été empêchées d'envoyer ou de recevoir des messages sur Facebook la veille d'une action en ligne prévue contre une entreprise d'investissement spécifique. Les groupes avaient déjà protesté en ligne contre la même entreprise.

Facebook a d'abord prétendu que les organisations avaient déjà commis une violation de la propriété intellectuelle, puis a affirmé qu'il s'agissait d'une erreur et a restauré les comptes progressivement, après la date de la manifestation prévue.[113]

Afin de remplir leurs obligations positives de garantir la liberté de réunion et d'association et, de fait, la liberté d'expression de manière plus générale, il est nécessaire que les États se dotent des pouvoirs juridiques requis et que les accords d'autorégulation et de corégulation soient conçus de manière à éviter la mise en œuvre de restrictions injustifiées. Des règles de recours et des sanctions appropriées et dissuasives devraient être mises en place pour décourager les intermédiaires d'internet de commettre des erreurs.

La liberté de réunion et d'association est déjà menacée par un environnement où la modération et la conservation du contenu sont assurées par des systèmes d'intelligence artificielle souvent opaques, ce qui signifie que l'"espace public" peut être différent pour chacun. Les points de référence communs sont rendus moins faciles à identifier, la formation de l'opinion étant en partie influencée par cette technologie.

[112] Comité des droits de l'homme des Nations Unies, Observation générale n° 37 (2020) sur le droit de réunion pacifique (article 21), septembre 2020, https://tbinternet.ohchr.org/_layouts/15/treatybodyexternal/Download.aspx?symbolno=CCPR%2fC%2fGC%2f37&Lang=en , consulté le 25 septembre 2020.

[113] Oliver Milman (2002) "Facebook suspend les groupes environnementaux malgré sa promesse de lutter contre la désinformation", *The Guardian*, www.theguardian.com/environment/2020/sep/22/facebook-climate-change-environment-groups-suspended , consulté le 8 octobre 2020.

Si la modération des contenus réduit délibérément ou accidentellement la diversité des informations dont disposent les individus ou les groupes, elle porte également atteinte à leur liberté de réunion et d'association. Il est en outre important de noter que les plus grands intermédiaires d'internet vendent désormais des services d'influence électorale microciblés. Cela soulève des questions importantes pour la démocratie et d'éventuels conflits d'intérêts.[114] Cela pourrait conduire à des situations où les intermédiaires d'internet pourraient être enclins à être plus indulgents à l'égard du discours politique qui provient d'un politicien qui paie pour un tel ciblage ou qui génère une controverse, un engagement et, par conséquent, des revenus.

Étude de cas : Droit, politique et autorégulation dans le cadre du référendum irlandais sur l'avortement

L'Irlande a organisé un référendum sur la légalisation de l'avortement le 25 mai 2018. L'Irlande a des règles étendues sur les dépenses des campagnes électorales, avec des règles différentes sur les dépenses liées au référendum. Les règles relatives au financement des campagnes référendaires se concentrent sur les groupes qui reçoivent des fonds, tandis que les groupes nominalement "autofinancés" ne sont pas soumis au même niveau d'examen.

Il n'existe pas de réglementation directe de la publicité en ligne, même s'il existe des règles sur les affiches de campagne et une interdiction de la publicité payante à la télévision et à la radio. Le gouvernement irlandais n'a pas mis à jour la loi pour changer cela (bien que les gouvernements successifs depuis 2008 aient prévu d'établir une commission électorale pour traiter ces problèmes et que ce processus avance rapidement au moment où nous écrivons ces lignes), ce qui laisse une disparité entre les règles de publicité en ligne et hors ligne.

Cette situation floue entre le traitement des groupes financés et autofinancés et entre les règles en ligne et hors ligne a finalement conduit à des actions unilatérales de la part des intermédiaires d'internet, qui sont des sociétés privées recevant de l'argent des groupes de campagne. Cette action a eu un impact difficile à évaluer sur les actions et le succès des groupes. Rétrospectivement, cela aurait pu être évité si l'État avait pris des mesures à l'avance pour s'assurer que les règles étaient claires et équitables.

Dans les mois qui ont précédé le référendum, Google et Facebook ont accepté des publicités payantes pour les campagnes du "oui" et du "non". Twitter, en revanche, a refusé toute publicité, restant fidèle à ses règles sur la publicité relative aux services médicaux et à ses règles sur la publicité politique.

Le mardi 7 mai 2018, Facebook a pris la décision d'autorégulation d'interdire toute publicité en relation avec le référendum provenant d'annonceurs basés en dehors de l'Irlande.

Le mercredi 8 mai 2018, Google a pris la décision d'autorégulation d'interdire toute publicité en relation avec le référendum, quel que soit l'organisme qui l'a financée.

Cet exemple illustre la nécessité de rendre compte de la 'curation' du contenu et, le cas échéant, de l'orientation de l'État. Il montre également que les décisions y relatives prises par les intermédiaires d'internet peuvent avoir des conséquences financières pour eux, tant positives que négatives.

[114] Zuiderveen Borgesius et al (2018) "Online Political Microtargeting : Promises and Threats for Democracy", Utrecht Law Review, Volume 14, Issue 1, www.ivir.nl/publicaties/download/UtrechtLawReview.pdf , consulté le 28 août 2020.

4. Droit de recours

En général, l'accent devrait être mis sur le fait d'éviter qu'il soit nécessaire de recourir à des mécanismes de recours pour des décisions incorrectes en matière d'autorégulation ou de co-régulation du contenu. Cela signifie qu'il faut s'assurer que toutes les mesures raisonnables sont prises par chaque partie prenante pour garantir que les droits ne sont pas violés. Lorsque des recours sont nécessaires, tous les outils nécessaires doivent être mis à disposition, tels que l'aide juridique et le soutien aux victimes, ainsi que l'accès à une procédure régulière.

L'équilibre des incitations pour les prestataires doit être tel que les recours et les informations sur la manière d'y accéder soient facilement accessibles.

a) Pour les victimes

Outre la cessation de l'infraction ou du délit, une approche d'autorégulation ou de corégulation de la modération de contenu ne peut pas faire grand-chose pour offrir une réparation aux victimes de comportements illégaux en ligne ou de délits hors ligne.[115] Il est donc essentiel que toute approche de la modération de contenu prévoie une coopération adéquate avec les autorités publiques et une action de leur part, afin de garantir la réparation des victimes - ce qui signifie qu'il faut apporter un soutien total pour annuler ou atténuer les dommages causés. Les obligations positives de l'État dans ce contexte ont été clairement énoncées par la Cour européenne des droits de l'homme dans l'affaire *K.U. c. la Finlande*, lorsqu'elle a rejeté l'idée que la réparation aurait pu être demandée à l'intermédiaire d'internet par les services duquel l'infraction en question avait été commise, comme alternative à la poursuite du contrevenant réel :

"*Il est évident que tant l'intérêt public que la protection des intérêts des victimes de crimes commis contre leur bien-être physique ou psychologique exigent l'existence d'un recours permettant d'identifier et de traduire en justice l'auteur réel de l'infraction, en l'occurrence la personne qui a placé l'annonce au nom du demandeur, et d'obtenir de lui une réparation financière*". [116]

b) Pour les personnes dont le contenu a été injustement supprimé

Il est essentiel que tout projet d'autorégulation ou de corégulation comporte des garanties adéquates pour éviter que le contenu ne soit suspendu ou retiré injustement. Conformément à la logique selon laquelle les droits sont la règle et les restrictions l'exception, le contenu devrait être laissé en ligne si possible et n'être suspendu ou retiré que dans des cas exceptionnels où cela est nécessaire. Ce qui est possible dépend en grande partie du type de contenu. S'il est nécessaire de retirer rapidement un contenu de l'Internet, cela devrait bien sûr être possible, tant que le processus reste prévisible, nécessaire et proportionné.

Cependant, il est inévitable que des erreurs se produisent. Garantir une transparence adéquate pour éviter que des erreurs similaires ne se reproduisent à l'avenir doit être considéré comme un élément fondamental de tout système de recours. Dans le même ordre d'idées, une transparence adéquate est essentielle pour identifier les éventuelles discriminations dans la manière dont les plaintes concernant différents types de contenus sont traitées. Les recours pourraient être décidés par les

[115] Cela dit, il convient de veiller à ce que la modération du contenu soit aussi sensible que possible aux besoins de ceux qui font des rapports dans de tels contextes. Ce point est largement discuté dans Brown Alexander, "Models of Governance of Online Hate Speech", étude indépendante réalisée pour le Conseil de l'Europe, pp. 155-167, https://rm.coe.int/models-of-governance-of-online-hate-speech/16809e671d , consulté le 18 janvier 2020.
[116] *K.U. c. Finlande*, no.2872/02, 2 decembre 2008, para. 47.

tribunaux mais aussi, si toutes les parties sont d'accord, par des mécanismes alternatifs de règlement des litiges ou des tribunaux électroniques.

Lorsque le contenu est retiré, il est également important que les mesures de transparence indiquent clairement les raisons spécifiques pour lesquelles le contenu a été retiré car, sans cela, il est difficile pour les individus de savoir si un recours est valable ou possible. Le droit à une procédure de recours rapide, accessible et équitable devrait toujours être accordé. Le droit à un recours effectif n'exige pas automatiquement une décision judiciaire (car les utilisateurs et les intermédiaires d'internet peuvent accepter de recourir à des mécanismes alternatifs de règlement des litiges), mais cela devrait toujours être une option.[117]

Il est simple, mais trompeur, de supposer que les coûts des erreurs de modération de contenu sont toujours facilement quantifiables en termes financiers. Les droits de l'homme sont inestimables, tout comme les valeurs démocratiques. Il est utile de réfléchir au préjudice causé par des restrictions injustifiées des droits de l'homme de cette manière plutôt que par le prisme d'une perte financière potentielle lorsque l'on envisage une réparation. Étant donné que le préjudice causé aux personnes dont le contenu est retiré injustement n'est souvent pas clairement financier, le fait de se baser sur de simples calculs des pertes possibles directement imputables aux retraits injustifiés ne permettra généralement pas d'indemniser les personnes concernées. Cela ne compensera pas non plus la société pour l'effet dissuasif de ces restrictions. En outre, ces coûts financiers normalement minimes n'inciteront pas les intermédiaires d'internet à faire preuve de plus de prudence à l'avenir. Dans le même temps, les intermédiaires d'internet devraient, à leur tour, avoir le droit d'obtenir une réparation significative et dissuasive contre les signalements négligents ou délibérément incorrects de contenus interdits par des personnes ou des organisations. Les sanctions pour les dommages causés par une modération excessive des contenus doivent donc tenir compte des préjudices plus larges causés à la société et doivent être dissuasives plutôt que simplement compensatoires.

Le rôle de l'État pour garantir que les restrictions des droits de l'homme liées aux entreprises soient corrigées a été clairement mis en évidence dans les principes directeurs des Nations unies sur les entreprises et les droits de l'homme :

> *"Si les États ne prennent pas les mesures appropriées pour enquêter sur les violations des droits de l'homme liées aux entreprises, les punir et les réparer lorsqu'elles se produisent, le devoir de protection de l'État peut être rendu faible ou même insignifiant. "*[118]

c) Adéquation du droit de recours et de réparation

Dans les cas où le contenu ou les services faisant l'objet de restrictions ont eu un impact négatif démontrable sur une partie importante du public, par exemple la diffusion d'une désinformation dangereuse, des [119]mécanismes devraient être mis en place pour remédier à cette situation en communiquant des messages équivalents à un nombre équivalent de personnes ou, si les données sont disponibles, aux mêmes personnes, dans le même format et à la même fréquence.[120] La

[117] Voir notamment le paragraphe 1.5.2 de la Recommandation CM/Rec(2018)2 du Comité des Ministres du Conseil de l'Europe sur les rôles et responsabilités des intermédiaires d'internet.

[118] Nations unies, "Principes directeurs de l'ONU sur les entreprises et les droits de l'homme", 2011, www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf , consulté le 2 septembre 2020.

[119] Le travail en cours du Groupe des régulateurs européens des services de médias audiovisuels est important dans ce contexte. Pour plus d'informations, voir https://erga-online.eu/ , consulté le 12 octobre 2020.

[120] Pour être efficace, un recours doit être capable d'apporter une réparation directe à la situation contestée - voir *Pine Valley Developments Ltd et autres c. Irlande*, décision de la Commission, no. 12742/87, 3 mai 1989.

publication d'une correction qui sera considérée par dix personnes comme une réparation pour un message vu par dix millions de personnes n'est pas une réponse adéquate.

IV Objectifs et moteurs de la modération du contenu

Ce chapitre explore les objectifs et les moteurs de la modération de contenu. En effet, lors de l'élaboration d'une politique en matière de modération de contenu, il est essentiel de tenir compte du fait que les raisons de la modération de contenu, ainsi que les types de contenu qui sont modérés, varient considérablement. Pour élaborer une politique efficace, il est essentiel de comprendre le type de contenu réglementé et les résultats escomptés.

Ce chapitre examine la modération de contenu sous deux angles, à savoir lorsqu'elle est mise en œuvre pour des raisons commerciales et pour des raisons de politique publique.

1. Modération du contenu et intérêts commerciaux

Afin de protéger adéquatement les droits de l'homme en ce qui concerne la modération du contenu, il est important de comprendre que, du point de vue des intermédiaires d'internet, cette activité a de multiples moteurs. Elle peut être entreprise :

- pour garantir que le modèle commercial reste valable - une plateforme axée sur les voitures peut souhaiter identifier et supprimer les contenus associés à des sujets autres que les voitures ;
- pour s'assurer qu'il n'y a pas de contenu sur la plateforme qui serait déplaisant pour les annonceurs et qui réduirait donc les revenus ;
- pour éviter d'être tenu responsable d'un contenu potentiellement illégal ou
- d'être perçu comme faisant un effort pour l'amélioration de la société en luttant contre les contenus potentiellement illégaux.

Un contenu problématique exacerbé par les modèles économiques

De nombreux intermédiaires d'internet ont un modèle commercial qui dépend partiellement ou totalement de la collecte et de l'utilisation des données personnelles des utilisateurs. Ces données sont générées lorsque les utilisateurs s'intéressent au contenu de la plateforme. La conception du service de l'intermédiaire d'internet peut servir à stimuler cet engagement (en récompensant les personnes qui publient un contenu qui stimule l'engagement avec des "j'aime" ou des commentaires positifs similaires).

En conséquence, les intermédiaires d'internet sont confrontés à un conflit d'intérêts potentiel si le contenu qui conduit à des niveaux d'engagement plus élevés est également un contenu illégal ou autrement indésirable. Une étude, bien qu'à petite échelle,[121] a révélé qu'à un moment donné, un quart des vidéos les plus regardées sur YouTube, sur Covid-19, contenaient des informations trompeuses.[122] Les conflits d'intérêts potentiels doivent être identifiés et atténués lors de la planification de tout système d'autorégulation ou de corégulation. Si le modèle commercial et les intérêts d'une plate-forme sont à l'origine d'un comportement illégal ou indésirable, c'est ce problème qu'il faut résoudre, et non la rapidité avec laquelle le comportement peut être supprimé ou rétrogradé. Bien que des exceptions puissent exister, un système d'autorégulation ou de corégulation

[121] Vidéos en anglais, non dupliquées, d'une durée inférieure à une heure.
[122] Li H.O., Bailey A., Huynh D. et al (2020) "YouTube as a source of information on COVID-19 : a pandemic of misinformation ?", *BMJ Global Health*, https://gh.bmj.com/content/5/5/e002604 , consulté le 20 mai 2020.

ne doit pas reposer sur des intermédiaires d'internet agissant de manière incompatible avec leurs propres intérêts financiers ou avec les fondements de leur modèle commercial.

Dans certains cas, les sociétés de médias sociaux sont également directement impliquées dans le suivi en ligne sur les sites d'information, par exemple. Ainsi, la diffusion d'une nouvelle sensationnelle ou trompeuse peut apporter à l'intermédiaire d'internet des revenus provenant à la fois des données que cette nouvelle génère à partir du site d'information d'origine et de l'engagement que cette nouvelle génère lorsqu'elle est publiée sur leur plateforme, créant ainsi une deuxième couche de conflit d'intérêts.

Le risque de conflits d'intérêts potentiels est d'autant plus grand lorsque des intérêts économiques supplémentaires entrent en jeu. De nombreux intermédiaires d'internet vendent également de l'influence économique et politique en microciblant les consommateurs et les électeurs.

En revanche, les plateformes de médias sociaux décentralisées, comme Mastodon, permettent aux petites et grandes communautés de définir et de gérer leurs propres politiques de modération, et donnent aux utilisateurs plus de choix sur les personnes qu'ils suivent dans ces communautés, que les plateformes centralisées.[123] Il est important de noter que la décentralisation contribue à minimiser l'un des problèmes majeurs du "marché" actuel, à savoir la modération du contenu à l'échelle. D'une part, Mastodon, étant décentralisé et open source, permet des espaces d'interaction sûrs et fortement réglementés mais, d'autre part, ne peut empêcher la mise en place d'espaces non réglementés. Toutefois, cela se fait sans risque d'amplification vers d'autres publics, comme cela tend à se produire dans les plateformes centralisées et axées sur les données. La Commission fédérale allemande des données a lancé sa propre instance de Mastodon en 2020.[124]

En conclusion, les incitations changeantes, et éventuellement contradictoires, des acteurs du marché, ainsi que les conséquences de la conception des services en ligne, doivent être pleinement comprises et prises en compte lorsque des projets d'autorégulation et de corégulation sont conçus ou demandés par les autorités publiques, afin de garantir la prévisibilité, la proportionnalité, la légitimité et l'efficacité.

L'étude de cas ci-dessus sur le référendum irlandais sur l'avortement montre qu'il n'y a parfois pas de décisions neutres, l'inaction même totale de l'État ayant un impact significatif sur le processus démocratique.

2. Modération du contenu pour des raisons de politique publique

Afin d'élaborer des politiques efficaces pour traiter les contenus problématiques en ligne, il est essentiel de comprendre la nature de ces contenus. Il est peu probable que des problèmes radicalement différents nécessitent des solutions identiques. Les catégories suivantes visent à illustrer l'éventail des contenus qui sont traités et ne sont pas censées être définitives ou immuables. Certains types de contenu peuvent être classés dans plusieurs catégories. Les États doivent veiller à ce que le rôle que les i intermédiaires d'internet sont censés jouer dans la lutte contre les contenus illicites en ligne soit adapté au type de contenu illicite en question.

[123] Pour une analyse plus approfondie des questions relatives à la décentralisation et à l'interopérabilité, voir Brown Ian, "Interoperability as a tool for competition regulation", 30 juillet 2020. https://osf.io/preprints/lawarxiv/fbvxd/ , consulté le 9 octobre 2020.

[124] Voir https://social.bund.de/@bfdi/105026921216079123 , consulté le 13 octobre 2020.

a) Un contenu qui est illégal partout, quel que soit le contexte

Il y a très peu d'exemples de contenus qui sont illégaux partout. L'exemple le plus clair est le matériel d'abus sexuel sur des enfants.[125] Les représentations d'abus d'enfants (parfois appelées "pornographie enfantine dans les textes anciens")[126] sont interdites dans la région du Conseil de l'Europe et au-delà par l'article 9 de la Convention de Budapest, l'article 20 de la Convention de Lanzarote et par plusieurs instruments juridiques internationaux, tels que la Convention 182 de l'Organisation internationale du travail, la Convention des Nations unies relative aux droits de l'enfant, et d'autres encore.

L'approche de ce problème en Europe (notification et retrait, utilisation de lignes directes, certains pays mettant en œuvre divers types de blocage du Web) a été presque statique au cours des quinze à vingt dernières années, sans évaluation significative de l'efficacité des mesures en vigueur ni de l'évolution de la criminalité. Les questions de savoir si, comment, pendant combien de temps et par qui, avec quels niveaux de transparence, les données doivent être stockées dans ce contexte commencent seulement à être discutées. Pour ces crimes, une évaluation continue est indispensable afin de garantir l'efficacité permanente des mesures prises pour les combattre.

b) Contenu illégal faisant partie d'un crime plus large

Ces infractions sont généralement plus graves et plus urgentes. Pour que ce matériel soit disponible en ligne, comme l'offre à la vente de produits issus d'une espèce animale protégée, il est probable qu'au moins un autre délit ait été commis. Toute initiative de modération du contenu qui ne prend pas en compte les éléments hors ligne d'un délit risque de laisser les victimes sans recours. En soi, un système d'autorégulation ou de corégulation ne peut rien faire pour enquêter sur les éléments hors ligne de ces infractions ou les sanctionner de manière dissuasive.

Il est donc crucial que toute initiative d'autorégulation ou de corégulation visant à lutter contre la criminalité grave (qui constitue une menace pour la vie humaine ou la maltraitance des enfants, par exemple) comprenne des responsabilités pour les États afin qu'ils prennent toutes les mesures nécessaires pour s'attaquer au problème plus large. Il ne devrait jamais être possible d'adopter une approche d'autorégulation ou de corégulation concernant ce type de contenu sans faire explicitement référence à l'engagement attendu avec les autorités répressives et les autres autorités étatiques compétentes. Par exemple, la loi allemande *sur l'amélioration de l'application de la loi dans les réseaux sociaux* (NetzDG) exige que certaines données associées soient stockées au cas où elles seraient nécessaires à des enquêtes ultérieures. Ces mesures doivent être calibrées très soigneusement afin d'éviter des conséquences involontaires sur la vie privée et d'autres droits de l'homme.

c) Contenu qui ne fait pas nécessairement partie d'une infraction plus large,

À l'autre bout du spectre, certaines infractions en ligne peuvent ne pas avoir de composante hors ligne. Par exemple, si quelqu'un télécharge une copie d'un film sur son site web ou sur un service de média social, le contenu lui-même n'est pas illégal, l'acte de télécharger le contenu peut ou non être illégal, peut être soumis à des exceptions et limitations légitimes du droit d'auteur et, en effet, si personne ne télécharge réellement le film, aucun préjudice réel n'a été subi par le(s) propriétaire(s) des droits sur le film. Par conséquent, une modération du contenu axée sur la disponibilité abordera

[125] Même dans ce cas, les lois ne sont pas toujours uniformes, les États disposant d'une certaine souplesse en ce qui concerne, par exemple, la législation nationale sur l'âge, l'âge apparent, la possession pour usage privé, la création et la possession par des enfants en vertu d'instruments tels que la Convention de Lanzarote (traité 201 du Conseil de l'Europe) et la directive européenne 2011/93 relative à la lutte contre les abus sexuels et l'exploitation sexuelle des enfants et la pédopornographie.

[126] Ce terme est tombé en désuétude et est désormais généralement évité.

cette question de manière plus complète que s'il s'agissait d'une infraction avec une composante hors ligne. Si la disponibilité est le seul problème, la suppression de la disponibilité le résout (bien qu'avec le risque de créer d'autres problèmes pour la liberté d'expression, le droit à la réparation, etc.). Inversement, si la disponibilité n'est pas le seul problème, la suppression de la disponibilité ne le résout pas.

d) Contenu légal qui est illégal principalement en raison de son contexte

Cela comprendrait, par exemple, la "pornographie de la vengeance" ou la publication non autorisée d'informations personnelles, lorsque le contenu lui-même n'est pas illégal, mais que la manière dont il devient disponible constitue une infraction.[127] Cela peut être difficile, voire impossible, à identifier sur la seule base d'une évaluation du contenu en question. D'autre part, le "doxing" sur les médias sociaux, par exemple, peut parfois être très facile à identifier.

e) Les contenus qui sont illégaux principalement en raison de leur intention.

Il s'agit par exemple de l'incitation à la violence ou de l'incitation au terrorisme. Ce ne sont pas les mots eux-mêmes, mais plutôt l'intention, le contenu et le statut de l'orateur qui conduisent à la commission de l'infraction.[128] [129] [130]

f) Contenu potentiellement préjudiciable mais pas nécessairement illégal

C'est une catégorie très large. Elle couvre, par exemple, l'aide ou les conseils juridiques, mais éventuellement dangereux, liés à l'automutilation ou au suicide. En fonction de leur modèle commercial ou de la clientèle visée, les intermédiaires d'internet peuvent choisir d'essayer de restreindre ces contenus. De telles restrictions ne peuvent pas porter atteinte à la liberté d'expression tant que les informations ou les idées "qui offensent, choquent ou perturbent l'État ou une partie de la population" disposent de moyens adéquats pour s'exprimer. Ce principe a été souligné par les sociétés informatiques et par la Commission européenne sur la première page de leur "Code de conduite pour lutter contre les discours de haine illégaux sur Internet" et constitue une reconnaissance par les sociétés participantes de leur rôle dans la défense des valeurs clés des droits de l'homme dans le contexte des systèmes de corégulation.[131]

La tâche, comme toujours, devrait être de définir le résultat le plus souhaitable, ce qui implique d'identifier le type de contenu, le public pour lequel il est considéré comme nuisible et la nature du préjudice redouté.

---

[127] Le Plan d'action de Rabat (www.ohchr.org/EN/Issues/FreedomOpinion/Articles19-20/Pages/Index.aspx , consulté le 18 janvier 2020) établit six critères sur "la liberté d'expression contre l'incitation à la haine" afin de fixer un "seuil élevé de restriction de la liberté d'expression". Ces critères sont les suivants : (1) Le contexte social et politique, (2) la position ou le statut de l'orateur, (3) l'intention d'inciter le public à s'opposer au groupe cible, (4) le contenu et la forme de la déclaration, (5) l'étendue de la diffusion, (6) la probabilité de préjudice, y compris l'imminence. Des critères similaires ont été utilisés par la Cour européenne des droits de l'homme, par exemple dans les affaires *Leroy c. France*, no. 36109/03, 2 octobre 2008, *Jersild c. Danemark*, no. 15890/89, 23 septembre 2004, *Feret c. Belgique*, no. 15615/07, 16 juillet 2009 et *Vejdeland et autres c. Suède*, no. 1813/07, 9 fevrier 2012.

[128] Ibid.

[129] Voir également L'Article 19 (2015) "Discours de haine expliqué" : A Toolkit", www.article19.org/data/files/medialibrary/38231/ "Hate-Speech'-Explained---A-Toolkit-%282015-Edition%29.pdf , consulté le 4 janvier 2021.

[130] Comité des Nations unies pour l'élimination de la discrimination raciale, Recommandation générale 35 sur la lutte contre les discours de haine raciste, 2013, p.5, www.refworld.org/docid/53f457db4.html , consulté le 19 janvier 2021.

[131] Code de conduite sur la lutte contre les discours haineux illégaux sur Internet, 30 juin 2016, https://ec.europa.eu/newsroom/just/document.cfm?doc_id=42985 , consulté le 21 mai 2020.

g) Un contenu qui suscite des préoccupations politiques

N'importe laquelle des catégories ci-dessus pourrait, à un moment donné, entrer dans cette catégorie. Il est dans la nature des médias que certains types de contenu deviennent soudainement le sujet de l'attention des médias, sans être nécessairement, en soi, un problème qui doit être, ou peut être efficacement, traité par la modération ou la réglementation du contenu. Par exemple, les "contrefaçons profondes" peuvent ne pas atteindre un niveau qui nécessite une intervention politique, mais un cas très médiatisé peut créer une impulsion pour agir. Sur le plan politique ou pour les besoins des relations publiques de l'entreprise, il est facile de se laisser prendre dans une spirale de réactions instinctives parce que "quelqu'un devrait faire quelque chose". Une approche claire et prévisible de l'élaboration de réponses proportionnées et efficaces devrait atténuer la pression de telles réactions spontanées, tout en améliorant la qualité de ces réponses.

3. Conclusion

Il est clair que ces types de contenu sont fondamentalement différents, non seulement en termes d'illégalité, mais aussi de caractéristiques et de gravité des conséquences. Il est peu probable qu'une violation du droit d'auteur soit traitée de la même manière qu'un téléchargement de matériel pédopornographique, pour au moins un type de contenu, voire les deux. Il est donc crucial d'adapter les réponses de modération de contenu au problème spécifique qu'elles tentent de résoudre.

Il convient également de noter, bien sûr, que les lois peuvent changer, de sorte qu'un contenu qui est légal dans un pays un jour peut devenir illégal le lendemain et vice versa.

V. Structures pour la modération du contenu

L'autorégulation et la corégulation sont des approches courantes et souvent très efficaces pour répondre aux préoccupations politiques de vastes pans de l'industrie. Toutefois, ces termes sont compris de différentes manières dans ces différents contextes et ont des impacts différents dans des environnements différents. Cette section examine les concepts d'autorégulation et de corégulation avant de tirer quelques conclusions de l'expérience concernant les caractéristiques communes que l'on peut trouver dans certaines approches réussies.

1. Autorégulation

L'autorégulation est la réglementation d'une entreprise ou d'un secteur d'activité en vue d'atteindre un objectif industriel ou de politique publique. Elle peut également être mise en œuvre en tant que stratégie visant à éviter la réglementation traditionnelle.

Les vastes succès de l'autorégulation dans le secteur des médias ont montré l'impact positif que cette approche peut avoir dans certaines circonstances. Lorsque l'autorégulation a bien fonctionné, elle a conduit à la création, par exemple, de codes d'éthique, de médiateurs et de mécanismes de plainte innovants qui permettent aux médias de rester indépendants tout en maintenant des normes élevées. En général, les entreprises de médias sont fortement incitées (par exemple, à maintenir des normes éditoriales élevées, à préserver leur réputation, leur flexibilité et leur indépendance) à assurer le succès de telles initiatives. Bien entendu, lorsque ces incitations n'existent pas, en raison de l'ingérence de l'État ou du manque d'indépendance vis-à-vis des influences politiques, elles sont moins présentes.

En principe, certains des mêmes avantages peuvent s'appliquer en ce qui concerne la modération du contenu en ligne, la flexibilité étant un avantage clé. Si un intermédiaire d'internet se réglemente lui-même (par exemple en choisissant d'autoriser ou non la publicité politique), s'il a un intérêt commercial évident à assurer l'efficacité (par exemple en filtrant la publicité non sollicitée par courrier électronique) et si l'effort a un avantage perçu pour ses utilisateurs, il peut bien fonctionner. L'inverse est également vrai, lorsque les incitations du fournisseur ne sont pas alignées sur les intérêts du régulateur ou des utilisateurs, l'autorégulation a moins de chances de bien fonctionner.

Parfois, les intermédiaires d'internet peuvent avoir des intérêts contradictoires, par exemple leur intérêt à protéger leurs utilisateurs contre la désinformation, mais aussi leur intérêt à tirer des revenus de la controverse et de l'engagement qui peuvent découler de la diffusion de ces types de contenu.

En outre, il est courant que les intermédiaires d'internet aient plusieurs modèles commerciaux fonctionnant en parallèle. Un fournisseur de médias sociaux peut ne tirer aucun revenu de sa fonction de médias sociaux, mais de la publicité, de la collecte de données sur les utilisateurs ou de la fusion de ces données avec des données provenant de tiers. Cela signifie que l'équilibre des incitations de l'entreprise pour assurer le succès de toute initiative de ce type peut être évolutif et imprévisible, même pour l'entreprise elle-même.

Le terme "autorégulation" ne doit être utilisé que pour désigner les situations dans lesquelles une entreprise ou un groupe d'entreprises agit pour réguler ses propres activités, sans pression directe ou indirecte de l'État. Il est important d'être clair sur la terminologie, car le degré d'implication de l'État est significatif pour les responsabilités juridiques de l'État.

Un intermédiaire d'internet peut être impliqué dans un large éventail de types d'autorégulation allant de la petite échelle (accepter ou non la publicité politique) à la grande échelle (contrôler toutes les expressions potentiellement nuisibles de ses utilisateurs). Cette autorégulation peut être intéressée (prévention du spam) ou non intéressée (prévention de contenus qui pourraient être rentables). Cela signifie que peu, voire pas du tout, de suppositions doivent être faites sur l'opportunité de l'autorégulation en ce qui concerne la modération de contenu par les intermédiaires d'internet.

2. La corégulation

Lorsque l'État et les acteurs privés coopèrent pour créer un cadre ad hoc afin de résoudre un problème de politique publique, on parle de "corégulation". Elle peut également couvrir les situations dans lesquelles les associations industrielles adoptent des codes dont l'adhésion peut être utilisée pour démontrer le respect des obligations légales.

La Cour européenne des droits de l'homme a déclaré que les mécanismes d'autorégulation et de corégulation peuvent être acceptables, à condition qu'ils comportent des garanties effectives des droits et des recours effectifs en cas de violation des droits.[132]Pour les mesures de corégulation, la Cour exige un degré considérable d'implication des gouvernements, comme l'approbation des règles.

Les mécanismes de corégulation devraient être basés sur un cadre juridique mis en place par l'État. Un tel cadre devrait définir des limites claires et prévoir des garanties pour empêcher les décisions arbitraires des agents non étatiques.[133]

[132] *Peck c. Royaume-Uni*, no. 44647/98,28 janvier 2003I, paras. 108 et 109.
[133] Kuczerawy Aleksandra (2016) "Private enforcement of public policy : freedom of expression in the era of online gatekeeping", thèse de doctorat, KU Leuven.

En général, les mêmes considérations s'appliquent à l'application des approches de corégulation qu'à celle des approches d'autorégulation. Toutefois, les approches de corégulation permettent à une autorité publique diligente d'exiger plus de transparence et de responsabilité que ce qui pourrait être le cas autrement.

Par exemple (et même si cette corégulation est appelée autorégulation par la Commission européenne), les Principes directeurs sur " L'approche 'suivre l'argent' pour le respect des droits de propriété intellectuelle" sont un bon exemple des types de garanties qui peuvent être envisagées.[134] Cet accord a été négocié entre la Commission européenne et les parties prenantes de l'industrie et de la société civile. Il promettait que le Protocole d'accord qui en découlerait développerait des indicateurs de performance clés, un processus de "vérification et de conformité" évalué de manière indépendante et qu'un équilibre approprié entre les différents droits fondamentaux en jeu serait assuré et démontré. Une telle approche permettrait la protection des droits de l'homme, la responsabilité, la flexibilité et une transparence significative, ainsi qu'une vérification indépendante que les objectifs de politique publique sont également atteints.[135]

Il convient toutefois de noter qu'en réalité, l'autorégulation et la corégulation n'existent généralement pas en tant que deux concepts clairs et distincts. Ces initiatives naissent souvent sous la pression des gouvernements, de la perception de l'industrie selon laquelle la législation est en suspens et doit être anticipée, ou de "menaces" manifestes de législation de la part des gouvernements. En ce qui concerne les projets "suivre l'argent " de la Commission européenne, celle-ci a annoncé en 2015 que la législation pourrait suivre si l'approche "autorégulatrice" n'était pas pleinement efficace. En outre, malgré l'appel, la convocation et la négociation du code conduite pour les annonceurs "suivre de l'argent " et malgré les "principes directeurs" du code donnant à la Commission des tâches spécifiques, la Commission européenne continue de qualifier cet instrument d' "autorégulation".[136]

Lorsque des priorités de politique publique sont en jeu, il ne semble guère utile d'adopter une approche qui soit purement autorégulatrice ou un mélange peu clair d'autorégulation et de corégulation. Une approche claire, ciblée et responsable, fondée sur un engagement constructif des autorités publiques, a plus de chances d'aboutir à des résultats positifs.

3. Caractéristiques communes des approches réussies

Il y a beaucoup à gagner à revoir l'histoire de l'autorégulation et de la corégulation, y compris dans d'autres secteurs, pour éviter les pièges et adopter des pratiques qui maximisent les chances de succès, tout en restant conscient de l'importance particulière des outils de communication en ligne pour les processus et institutions démocratiques, la paix et la stabilité, l'égalité et la non-discrimination.

Une étude a comparé les caractéristiques et les succès et échecs relatifs des régimes d'autorégulation dans les secteurs de la sylviculture, de la pêche, du tabac, des boissons non alcoolisées et de la

[134] Commission européenne, direction générale du marché intérieur, de l'industrie, de l'entrepreneuriat et des PME, principes directeurs «Le suivre de l'argent approche pour le respect des DPI - accord volontaire des parties prenantes sur la publicité en ligne et les DPI», https://ec.europa.eu/docsroom/documents/19462 , consulté le 12 mai 2020.

[135] Quatre ans après l'adoption des principes directeurs, la Commission européenne n'est pas en mesure de dire si ces engagements ont effectivement été respectés. Voir la question parlementaire 3115/2020 du député européen Patrick Breyer à la Commission européenne, www.europarl.europa.eu/doceo/document/E-9-2020-003115_EN.html , consulté le 1 septembre 2020.

[136] Voir par exemple https://ec.europa.eu/growth/industry/policy/intellectual-property/enforcement/memorandum-of-understanding-online-advertising-ipr_en , consulté le 12 mai 2020.

restauration rapide.[137] En examinant les succès relatifs des différents régimes, les chercheurs ont déterminé que la motivation derrière les initiatives peut être déterminante :

> *"Le type de motivation peut être un facteur déterminant de la réussite. Dans certains cas, une industrie perçoit qu'elle doit se surveiller elle-même parce que les gouvernements s'impliquent trop peu, comme ce fut le cas pour la gestion des forêts et des pêches. Pour d'autres industries, l'intervention du gouvernement est perçue comme une menace, et les actions d'autorégulation sont un moyen de prévenir ou d'anticiper une réglementation extérieure".*

La première motivation, un besoin évident des entreprises, en l'absence de leadership gouvernemental, semble être un facteur déterminant de la réussite de tous les programmes examinés. Cependant, l'autorégulation comme moyen de prévenir l'intervention du gouvernement a été un facteur d'échec. Cela semble logique, car l'objectif direct de l'initiative d'autorégulation est de prévenir la réglementation et non de résoudre le problème de politique publique lui-même.

Sur la base de l'expérience des différentes initiatives d'autorégulation examinées, les auteurs énumèrent neuf normes clés sur lesquelles l'autorégulation future dans le secteur alimentaire pourrait être construite, en tirant les leçons des erreurs et des faux pas du passé.

Objectif	Standard
Transparence	1) Des normes d'autorégulation transparentes créées par une combinaison de scientifiques (non rémunérés par l'industrie) et de représentants des principales organisations non gouvernementales, des parties impliquées dans la gouvernance mondiale (par exemple, l'Organisation mondiale de la santé, l'Organisation des Nations unies pour l'alimentation et l'agriculture) et de l'industrie
	2) Aucun parti ne se voit attribuer un pouvoir ou une autorité de vote disproportionnés
Des objectifs et des repères significatifs	3) Des codes spécifiques de comportements acceptables basés sur des critères scientifiquement justifiés
	4) Des critères de référence prédéfinis pour garantir le succès de l'autorégulation
Responsabilité et évaluation objective	5) Rapports publics obligatoires sur le respect des codes, y compris sur les progrès réalisés en matière de respect intégral des engagements et de respect des principaux critères de référence
	6) Des procédures intégrées et transparentes permettant à des tiers d'enregistrer des objections aux normes d'autorégulation ou à leur application
	7) Évaluation objective des critères d'autorégulation par des groupes externes crédibles non financés par l'industrie pour évaluer les résultats sanitaires, économiques et sociaux
	8) Évaluations et audits périodiques pour déterminer la conformité et les résultats

[137] Sharma L. L., Teret S. P., Brownell K. D. (2010) "L'industrie alimentaire et l'autorégulation : des normes pour promouvoir le succès et éviter les échecs en matière de santé publique", *American journal of public health*, 100(2), 240-246, 2011, https://doi.org/10.2105/AJPH.2009.160960 , consulté le 8 octobre 2020.

Objectif	Standard
Surveillance	9) Surveillance éventuelle par un organisme mondial de réglementation ou de santé approprié (par exemple, l'Organisation mondiale de la santé)

Une analyse similaire des initiatives d'autorégulation en matière de modération du contenu et d'autorégulation et de corégulation dans le secteur de l'internet semble très tardive et susceptible de générer des résultats similaires. Cette évaluation recoupe largement celle du Conseil national des consommateurs du Royaume-Uni. Ce dernier a constaté que les initiatives d'autorégulation doivent avoir des objectifs politiques clairs, ne doivent pas entraver les possibilités de concurrence, doivent comporter un élément indépendant fort tant pour la conception que pour la gouvernance, doivent avoir une structure institutionnelle spécifique et doivent fonctionner dans un cadre juridique clair.[138]

VI. Transparence

Pourquoi la transparence est essentielle

La transparence a deux composantes, la première concerne ce qui est fait ou tenté (à savoir chercher à imposer des « conditions de service » claires ou une loi spécifique). Le second, qui fait l'objet de la présente section, concerne les effets de ce qui est fait - détails sur ce qui a été retiré, combien a été retiré pour quels motifs, combien a été remis, le nombre de plaintes reçues concernant des contenus illicites ou des enlèvements, l'impact sur le(s) problème(s) traité(s), etc. Dans la mesure du possible, ces données doivent être produites selon une méthodologie standardisée et dans des formats lisibles par machine.

En particulier dans un environnement très fluide où tant les crimes que les technologies changent continuellement, et où la frontière entre les actions de l'État et les actions privées est souvent floue, il est impossible de garantir la prévisibilité, la nécessité et la proportionnalité en permanence sans les données permettant de réaliser ces évaluations.

Le coût de l'autorégulation et de la corégulation est souvent une réduction de la responsabilité et de la légitimité démocratique. Le principal avantage des approches d'autorégulation et de corégulation en général est qu'elles sont flexibles. Cet avantage est particulièrement précieux dans l'environnement en ligne en constante évolution. Les problèmes évoluant, les réponses, logiquement, doivent également changer. Toutefois, sans une transparence significative, il n'est pas possible d'identifier et d'évaluer les changements ni de procéder aux ajustements correspondants. Par conséquent, sans transparence, la société perd l'un des principaux avantages de l'autorégulation et de la corégulation, tout en en supportant le coût.

Pour garantir que les restrictions soient nécessaires et proportionnées

Les restrictions en matière de droits de l'homme doivent être nécessaires et proportionnées au moment où elles sont lancées, mais aussi de manière continue. En Belgique, un code de pratique a été

[138] Chris Jay Hoofnagle (2016) "Federal Trade Commission Privacy Law and Policy– Chapter 6 Online privacy " Cambridge University Press, UC Berkeley Public Law Research Paper No. 2800276, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2800276 , consulté le 28 mai 2020.

signé en vertu duquel les "newsgroups"[139] individuels sont retirés des serveurs d'un intermédiaire d'internet participant, si un groupe d'intérêt particulier demande leur suppression en raison d'un partage prétendument illégal de matériel protégé par des droits d'auteur.[140] En réponse, les utilisateurs contrôlaient au jour le jour quels groupes de discussion étaient disponibles sur les services des intermédiaires d'internet participants. Ils le faisaient pour identifier les groupes de discussion qui avaient été supprimés et, par extension, ceux qui étaient "les meilleurs" pour trouver des contenus non autorisés. Ils ont ensuite publié les listes des groupes de discussion en ligne, ce qui leur a permis, ainsi qu'à d'autres, de faire appel à d'autres fournisseurs pour accéder au contenu en question. En l'espace de quelques jours, le projet était devenu contre-productif.[141]

Il était très clair à l'époque que l'efficacité de la mesure s'est effondrée dans les quelques jours qui se sont écoulés entre son lancement et le moment où les utilisateurs ont modifié leur comportement en réaction. Dans la plupart des cas, ces évolutions seront beaucoup moins visibles et moins soudaines. En l'absence de mécanismes permettant d'identifier ces évolutions et d'y remédier en permanence, l'efficacité, la nécessité et la proportionnalité dans le temps ne peuvent être garanties. En particulier dans les situations où une action urgente est nécessaire, il faut veiller à ce que l'évolution des problèmes en question soit suivie avec soin.

Pour garantir la non-discrimination

La discrimination peut même se produire dans des systèmes de modération de contenu bien conçus et bien intentionnés. Elle peut se produire en raison du manque de ressources locales, de l'absence de connaissances sur l'utilisation régionale de la langue ou de l'utilisation différente des mots par différents groupes sociaux, du "jeu" (manipulation ciblée) de systèmes automatisés par de mauvais acteurs, et de nombreux autres facteurs. Il est donc essentiel que les systèmes de transparence soient conçus de manière à produire les données nécessaires pour identifier rapidement ces discriminations. Les systèmes qui ne disposent pas des mécanismes nécessaires pour produire les données permettant d'identifier une discrimination potentielle ne devraient pas être autorisés.

Pour garantir la responsabilité des parties prenantes (comme les États)

La transparence est également essentielle pour garantir la responsabilité au-delà des parties prenantes directes (le plaignant, l'intermédiaire d'internet et la personne ou le groupe dont le contenu fait l'objet de la plainte). C'est le cas le plus évident lorsque le contenu fait partie d'un délit hors ligne. Si des preuves d'un délit ont été identifiées, les données sur la rapidité avec laquelle elles ont été supprimées par l'intermédiaire d'internet donnent une image incomplète. En revanche, il faut des données sur le nombre de signalements qui ont été mis à la disposition des autorités répressives ou qui leur ont été transmis par celles-ci, ainsi que sur le nombre d'enquêtes et de poursuites qui ont été engagées à la suite de ces signalements.

Par exemple, les principaux intermédiaires d'internet utilisent une technologie photoADN à source fermée pour bloquer les téléchargements de matériel pédopornographique connu. Si des personnes téléchargent des contenus pédopornographiques connus, il convient de recueillir des données sur l'évolution de l'infraction en réponse à la mesure, si les autorités répressives ont la possibilité de prendre connaissance des cas de blocage, si les autorités répressives nationales accèdent aux données

[139] Des forums de discussion mondiaux qui sont reproduits, hébergés et mis à disposition par les fournisseurs de services pour leurs utilisateurs. Ces forums étaient populaires aux débuts de l'internet, mais ils sont aujourd'hui passés de mode.
[140] De Neeve Mick, "Belgian Providers Delete Newsgroups", Tweakers.net, 16 juillet 2005, https://tweakers.net/nieuws/38089/belgische-providers-schrappen-nieuwsgroepen.html , consulté le 12 mai 2020.
[141] Voir, par exemple, https://userbase.be/forum/viewtopic.php?f=24&t=7895&start=40 , consulté accès le 28 mai 2020.

personnelles associées et à quelle fréquence, si les données sont automatiquement mises à la disposition des autorités répressives ou non, et si cela conduit à des enquêtes ou des poursuites, afin de mesurer le succès de la méthode. En raison de la gravité du délit en question, des tests indépendants devraient être exigés. La suppression ou le blocage, en tant que stratégie autonome, crée essentiellement l'impunité pour les crimes graves.

Identification des données de transparence

Des efforts rigoureux et continus sont nécessaires pour identifier les données de transparence nécessaires. Il est essentiel de collecter toutes les données nécessaires pour mesurer le succès ou l'échec et pour identifier les incidences contre-productives ou négatives. Les critères de collecte des données doivent également être revus afin de s'assurer qu'ils ne créent pas d'incitations perverses (comme l'incitation à la rapidité plutôt qu'à l'exactitude, qui peut également entraîner des tentatives à « déjouer » des mécanismes de plainte au détriment des groupes vulnérables). Les données pertinentes qui devraient être collectées peuvent également être identifiées en examinant les risques possibles pour les droits de l'homme et les objectifs et cibles minimums de la modération du contenu lui-même.

Dans l'exemple ci-dessus, les mesures de lutte contre la maltraitance des enfants, il devrait être évident que les données nécessaires pour des raisons de transparence incluraient la raison spécifique du retrait du contenu (conditions de service ou loi), si les données sont stockées en relation avec du matériel potentiellement criminel, si les autorités répressives y ont accès sur demande, si ou combien de fois ces données ont été demandées par les autorités répressives, combien de fois le contenu n'a pas été immédiatement retiré afin d'éviter d'interférer avec les enquêtes des autorités répressives, d'autres raisons de retard dans les retraits, la rapidité des retraits/blocs, etc.

Reconnaître les incitations positives et négatives créées par les mesures de transparence

Si les gouvernements font pression sur les intermédiaires d'internet pour qu'ils prennent des mesures qu'ils n'auraient pas prises autrement, alors ces mesures seront motivées par les objectifs implicites ou explicites qui ont été fixés par les demandes des gouvernements. Si les paramètres clés sont "combien" et "à quelle vitesse", les fournisseurs sont incités à supprimer autant que possible, le plus rapidement possible.

Les exigences de transparence sans granularité appropriée peuvent conduire à la production de données impossibles à utiliser, que ce soit délibérément ou accidentellement. Par exemple, le rapport de transparence de YouTube en [142]vertu de la loi allemande sur l'application des réseaux comprend une catégorie appelée "contenu terroriste ou anticonstitutionnel" qui mélange le contenu terroriste avec des infractions aux dispositions du code pénal allemand généralement, mais pas totalement, sans rapport avec le terrorisme, comme l'utilisation de symboles d'organisations anticonstitutionnelles (article 86a du code pénal) et certains types de falsification (article 269 du code pénal). Le signalement de Facebook,[143] en vertu des mêmes dispositions de la même loi, est beaucoup plus granulaire. Il faut s'en féliciter.

Les divergences méthodologiques laissent l'État allemand et le peuple allemand sans données claires qui pourraient être comparées entre les intermédiaires d'internet et dans le temps, alors que c'est probablement la principale raison pour laquelle il y a des obligations de transparence au départ. Étant

[142] Voir https://transparencyreport.google.com/netzdg/youtube?hl=en

[143] https://about.fb.com/wp-content/uploads/2020/01/facebook_netzdg_January_2020_english.pdf , consulté le 13 mai 2020.

donné que le contenu est apparemment supprimé sur la base de dispositions spécifiques de la même législation, il semble contre-intuitif que les entreprises fournissent différents types et détails de données. Pour une transparence totale et efficace, les données devraient être fournies avec des niveaux de granularité maximums et une méthodologie identique, permettant une analyse et une évaluation efficaces des méthodes de modération du contenu appliquées. À titre d'exemple de bonne pratique en matière de transparence qui devrait être promue et reproduite, la ligne téléphonique d'urgence autrichienne pour les documents relatifs aux abus sur les enfants et au soutien/à l'adhésion à l'idéologie nationale-socialiste (stopline.at) publie des rapports de transparence avec une méthodologie rigoureusement cohérente. Cela donne aux décideurs politiques et autres les données nécessaires pour voir l'ampleur de problèmes particuliers à un moment donné et pour évaluer l'évolution des problèmes dans le temps.

Enfin, la transparence est également nécessaire dans la promulgation et l'adaptation des règles de contenu. Les modifications des règles de contenu doivent être faciles à comprendre et doivent être justifiées. De plus, il faudrait expliquer clairement pourquoi certains types de contenus qui ne sont pas illégaux sont interdits sur les services des intermédiaires d'internet.

Des signaleurs de confiance

Les intermédiaires d'internet trouvent souvent utile d'utiliser des organisations spécialisées comme filtre pour obtenir des rapports plus fiables sur les contenus illicites. En effet, le rôle des signaleurs de confiance est institutionnalisé dans la loi allemande sur l'application des réseaux. Ils sont appelés "signalisateurs de confiance" ou "signalisateurs prioritaires".[144] Des règles de transparence spécifiques sont nécessaires pour de telles initiatives afin de garantir qu'aucun conflit d'intérêt (structurel (travailler sur le contenu qu'il tente d'éliminer) ou financier (être financé par la plate-forme)) ne soit accidentellement créé et que leur niveau d'efficacité et de fiabilité reste constamment élevé. Le recours à des signaleurs de confiance ne devrait pas être obligatoire et les avis de ces derniers ne devraient pas être considérés comme une "connaissance réelle" de l'illégalité du contenu. Cela leur conférerait une fonction quasi judiciaire et pourrait entraver l'utilisation de cette approche.

Curieusement, les signaleurs de confiance n'existent que pour la restriction des contenus et non pour leur protection ou leur remise en ligne. Si l'on peut faire confiance à des tiers d'une manière qui amène les intermédiaires d'internet à donner la priorité à leurs notifications de contenu illégal et à les supprimer, il semble logique que les intermédiaires d'internet puissent également faire confiance à des tiers pour qu'ils soumettent des notifications de priorité indiquant que certains contenus ou comptes n'auraient pas dû être restreints.

Allant plus loin dans cette logique, les radiodiffuseurs de service public européens affirment que, étant donné qu'ils sont soumis à une responsabilité éditoriale directe et complète, les intermédiaires d'internet ne devraient pas les soumettre "à une forme quelconque de contrôle ou d'ingérence".[145] En d'autres termes, ils devraient être exclus de manière permanente du système du signalement de fiabilité ou de confiance . Cela soulève certaines questions fondamentales, notamment le degré d'indépendance nécessaire à un radiodiffuseur pour bénéficier de ce statut, et qui serait chargé de

[144] EuroISPA (2019) "Priority Flagging Partnerships in Practice", www.euroispa.org/wp-content/uploads/Hutty_Schubert_Sanna_Deadman-Priority-Flagging-Partnerships-in-Practice-EuroISPA-2019.pdf , consulté le 28 mai 2020.

[145] Réponse de l'Union européenne de radio-télévision à la consultation de la Commission européenne sur la loi sur les services numériques (question 16), www.ebu.ch/files/live/sites/ebu/files/Publications/Position_Papers/open/EBU_response_Digital_Services_Act_consultation%2008092020.pdf , consulté le 30 septembre 2020.

prendre ou de revoir la décision d'accorder ce statut à un radiodiffuseur? En outre, on pourrait soutenir que la même logique pourrait être utilisée pour dire que tout individu qui ne parle pas anonymement est soumis à la loi de son pays, ce qui pourrait entraîner une discrimination à l'encontre de ceux qui souhaitent ou doivent parler anonymement.

VII. Principes clés pour une approche de la modération de contenu fondée sur les droits de l'homme

1. Transparence

Comme expliqué ci-dessus, la transparence est l'élément le plus important pour parvenir à une modération réussie du contenu. Elle 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. Tous les critères énumérés ci-dessous reposent, dans une plus ou moins large mesure, sur la transparence pour être réalisés.

Des normes minimales doivent être définies pour évaluer si la modération du contenu en question atteint ses objectifs spécifiques. Il peut s'agir de normes pour les faux négatifs, les faux positifs et les temps de réponse. Cela signifie, par exemple, que des normes minimales devraient être fixées pour le nombre de fois où un contenu illicite est incorrectement étiqueté comme non illicite et où un contenu non illicite est incorrectement étiqueté comme illicite, avec des normes clairement définies pour les taux d'erreur acceptables. Cela nécessite un examen indépendant d'au moins un échantillon représentatif de cas. Tout dépassement des taux d'erreur acceptables devrait automatiquement donner lieu à des mesures correctives.

2. Les droits de l'homme par défaut

En vertu de la Convention, les droits de l'homme sont la norme et des restrictions peuvent être imposées exceptionnellement lorsque cela est nécessaire et proportionné. Cette approche doit guider l'élaboration des politiques en matière de modération du contenu.

Il est également important d'identifier de manière proactive les droits qui pourraient être menacés avant de lancer un processus de modération des contenus, tout en gardant à l'esprit que le fait de ne pas modérer les contenus peut nuire à l'égalité. La modération de contenu peut, par exemple, restreindre les droits protégés par les articles 8 et 10 de la Convention. Il s'agit de droits fondamentaux qui sont essentiels pour qu'une société démocratique puisse exister et prospérer. Le droit à un recours effectif, inscrit à l'article 13 de la Convention, tant pour les victimes d'infractions qui ont eu lieu entièrement ou partiellement en ligne, que pour celles dont les droits fondamentaux ont été restreints par des mesures de modération de contenu, doit être rigoureusement protégé.

En raison de l'évolution constante des crimes et des technologies, un examen préalable des mesures d'autorégulation ou de corégulation ne suffit pas pour garantir le respect des droits de l'homme. Un examen fréquent de l'impact ou des impacts des mesures est également essentiel. Les obligations positives et négatives des États en matière de protection des droits de l'homme sont également applicables dans l'environnement en ligne.

3. Identification des problèmes et des cibles

La modération de contenu constitue un effort pour résoudre un problème. Il est donc crucial que le problème soit identifié aussi clairement que possible, afin de trouver des solutions ciblées à des problèmes variés. Cela est important pour garantir la nécessité et la proportionnalité.

Il faut comprendre la nature du problème. La législation qui confie la charge de la gestion des risques (qui est, par définition, différente pour chacun) aux intermédiaires d'internet comporte des défis fondamentalement différents par rapport au retrait des contenus illicites. Le risque n'est pas, presque par définition, un préjudice.[146] Il est donc crucial, pour éviter des conséquences involontaires, que toute intervention politique ayant pour but de minimiser le risque soit clairement reconnue comme telle, 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 protection significative de la liberté d'expression, ainsi que d'outils permettant d'identifier les effets contre-productifs.

Si la modération du contenu est effectuée dans le cadre d'un système d'autorégulation ou de co-régulation, il faut également prévoir des mécanismes permettant de reconcevoir, d'adapter ou d'abandonner le projet, si les normes minimales ne sont pas respectées ou si la nature du problème évolue de telle sorte que l'approche identifiée n'est pas efficace.

4. Une décentralisation significative

Comme l'a expliqué David Kaye, rapporteur spécial des Nations unies sur la liberté d'opinion et d'expression, dans son essai sur ce sujet,[147] la décentralisation est nécessaire pour modérer le contenu dans des contextes multinationaux ou mondiaux. Une modération décentralisée, multipartite, rémunérée, responsabilisée et indépendante est essentielle pour traiter les problèmes au niveau régional, en tenant compte des particularités régionales lorsqu'il s'agit de traiter les types de contenu les plus difficiles.

> *Partout où les entreprises sont présentes sur le marché, elles devraient mettre en place des conseils multipartites, dont elles rémunéreraient les membres, pour les aider à évaluer les problèmes de contenu les plus difficiles, à évaluer les questions émergentes et à s'opposer aux plus hauts niveaux de direction de l'entreprise.*[148]

En outre, son analyse est que la clarté en matière de prise de décision algorithmique permettrait aux particuliers et aux universitaires d'"enregistrer des contestations sérieuses" de l'exécution des décisions. En l'absence de données facilement disponibles, la recherche et les contestations des décisions deviennent impossibles. Conformément aux exigences de transparence ci-dessus, des données adéquates doivent être mises à la disposition de la société civile et des chercheurs techniques et universitaires pour faciliter une analyse continue.

---

[146] Livingstone Sonia, Kalmus Veronika, Talves Kairi (2014) *Expériences des filles et des garçons en matière de risques et de sécurité en ligne*, dans : Carter Cynthia, Steiner Linda, McLaughlin Lisa (Ed.), The Routledge Companion to Media and Gender (190-200), Londres : Routledge, p. 192.

[147] David Kaye (2019) "A New Constitution for Content Moderation", medium.com, https://onezero.medium.com/a-new-constitution-for-content-moderation-6249af611bdf , consulté le 13 mai 2020.

[148] Idem.

5. Communication avec l'utilisateur

La modération du contenu implique une restriction des libertés fondamentales. Ces restrictions doivent respecter les normes des droits de l'homme et être aussi transparentes que possible envers le public, les plaignants, les victimes et ceux dont le contenu est retiré.

a) Clarté et accessibilité des conditions de service

Outre le respect total des droits de l'homme de toutes les parties prenantes, tous les outils disponibles doivent être utilisés pour garantir que les conditions de service d'un intermédiaire d'internet soient aussi claires et accessibles que possible. L'application de ces règles devrait également être prévisible, conformément au droit relatif aux droits de l'homme. Selon l'analyse de David Kaye, les normes relatives aux droits de l'homme ont développé un langage pour définir des cadres qui peuvent être utilisés afin d'articuler et d'assurer le respect des normes démocratiques, ainsi que contrer les demandes autoritaires. Ce langage doit être utilisé.

b) Clarté sur la communication avec les utilisateurs

Les personnes qui souhaitent se plaindre d'un contenu apparemment illégal, ou d'un contenu qui enfreint en apparence les règles internes d'un intermédiaire d'internet, doivent disposer des outils nécessaires pour communiquer leur plainte à l'entreprise de la manière la plus spécifique possible. Le cas échéant, les intermédiaires d'internet, dans le cadre d'un dialogue ouvert avec les organisations représentatives appropriées, devraient veiller à ce que leurs mécanismes de plainte tiennent compte des besoins des victimes.[149]

Ceux qui publient du contenu devraient bénéficier de règles claires, équilibrées et compatibles avec les droits de l'homme, mises en œuvre et appliquées de manière équilibrée et prévisible, et non à la discrétion de la plateforme. Ceux qui accèdent à ces contenus devraient également avoir un accès facilité à ces règles, ainsi que le droit et les outils nécessaires pour déposer des plaintes spécifiques, s'ils le souhaitent.

Le contenu ne doit pas être mis hors ligne immédiatement, s'il n'est pas urgent de le faire. Au contraire, la personne qui a téléchargé le contenu doit recevoir des informations claires sur les raisons pour lesquelles son contenu peut avoir enfreint les conditions de service ou la loi, avoir le droit de défendre son téléchargement dans un délai déterminé et, dans tous les cas, le droit à un recours utile.

Certains contenus doivent être mis hors ligne le plus rapidement possible, en raison de la nature du contenu ou de son impact sur les victimes. Ces contenus doivent être bien définis et le processus de révision, de suppression et, le cas échéant, de remise en ligne doit être prévisible, responsable et proportionné.

6. Des garanties administratives de haut niveau

a) Un cadre juridique et opérationnel clair

Un cadre juridique clair et prévisible est essentiel pour garantir que les restrictions sont prévues par des instruments qui ont force de loi ou qui ont la qualité de loi. Les États devraient veiller à ce que les conditions de service soient claires, équilibrées et appliquées de manière équitable. Les lois sur les contenus illicites devraient être aussi claires et harmonisées que possible au niveau international. Les

[149] Brown Alexander (2020), op cit., chapitre VII.

États devraient garantir l'existence d'autorités compétentes et indépendantes ayant le droit d'émettre des ordres de retrait. Enfin, il convient de veiller à ce que les intermédiaires d'internet ne soient pas indûment incités à restreindre la liberté d'expression ou d'autres droits de l'homme. Des règles de recours appropriées et dissuasives devraient être mises en place afin de décourager les signalements malveillants par les particuliers et de dissuader les intermédiaires d'internet de se conformer de manière excessive.

b) Supervision pour assurer le respect des droits de l'homme

Une transparence significative sur la gouvernance, les processus décisionnels et les détails sur la manière, le moment, le pourquoi et le combien, le contenu qui a été retiré ou non et pour quelle raison, peuvent constituer la base de mesures significatives pour identifier les violations des droits de l'homme. Toutes les données qui ne sont pas personnelles devraient être rendues publiques, sur la base de normes industrielles convenues sur la méthodologie et le format de ces rapports. Toutes ces données devraient être mises à la disposition des structures de gouvernance appropriées, décentralisées, multipartites et indépendantes sur le plan organisationnel.

Des garanties particulières sont nécessaires en matière de protection des données en ce qui concerne la modération du contenu. Il convient de faire preuve de prudence lors du traitement des données personnelles des plaignants et des personnes accusées de télécharger des contenus illégaux ou indésirables (contenus qu'un intermédiaires d'internet peut souhaiter restreindre et qui ne sont pas illégaux en *soi*). C'est notamment le cas lorsqu'une personne est accusée de poster des informations potentiellement illégales, lorsque ce contenu révèle des données personnelles sensibles et/ou lorsque cela conduit à conserver le traitement de données qui ne seraient autrement pas nécessaires à la fourniture du service.

Les États devraient également prendre toutes les mesures nécessaires pour identifier et prévenir les cas de surconformité et de discrimination en matière de modération de contenu.

c) Évaluation et atténuation du "jeu" des mécanismes de plainte

Un bon rapport sur la transparence permettra aux entreprises et au public d'identifier le jeu des mécanismes de plainte des entreprises. Cela devrait être considéré comme une tâche prioritaire et permanente pour les systèmes d'autorégulation et de corégulation, notamment parce que les informations disponibles indiquent que ces jeux sont particulièrement préjudiciables aux femmes et aux minorités. On peut le constater, par exemple, dans les cas de "Women on Waves" et "Kick Out Zwarte Piet" aux Pays-Bas, qui sont décrits ci-dessus.

d) Garantir la cohérence et l'indépendance des mécanismes de contrôle

Les données sont un élément essentiel pour garantir la cohérence et l'indépendance d'un mécanisme de contrôle. Si suffisamment de données sur les décisions sont rendues publiques et, si nécessaire, mises à la disposition de tiers indépendants et si suffisamment d'échantillons de cas sont mis à la disposition d'un organe indépendant et impartial pour un examen proactif, dont les conclusions sont prises en compte de manière significative par l'intermédiaire d'internet, alors un haut degré de cohérence et d'indépendance peut être assuré.

Il faut veiller tout particulièrement à ce que les utilisateurs se sentent responsabilisés et écoutés lorsque des recours sont introduits à propos d'un contenu qui est ou va être supprimé. De même, la transparence est nécessaire pour garantir que les plaignants reçoivent des informations adéquates

leur permettant de comprendre si et pourquoi leurs plaintes n'ont pas conduit à la suppression du contenu en question.

e) Reconnaître les défis humains de la modération du contenu

Outre les dispositions relatives à la gestion du personnel de la recommandation du Comité des Ministres sur les rôles et responsabilités des intermédiaires d'internet,[150] il convient également d'accorder l'attention nécessaire aux droits du travail et à la santé mentale de tous les travailleurs impliqués dans l'examen manuel de contenus qui peuvent être choquants, dérangeants ou susceptibles d'avoir un impact psychologique sur les personnes concernées. C'est notamment le cas lorsque les intermédiaires d'internet confient cette tâche à des tiers, éventuellement basés dans d'autres pays dont le droit du travail est différent et peut-être moins protecteur.

f) Garantir la protection de la vie privée et des données

La modération de contenu implique le traitement de quantités importantes de données personnelles relatives à l'utilisateur, au plaignant et à la nature du contenu en question. Il serait utile que les États membres du Conseil de l'Europe s'assurent que des bases juridiques adéquates existent dans le droit national pour ce traitement.

Les intermédiaires d'internet doivent également veiller à ce que la modération du contenu n'entraîne pas de violations involontaires de la protection des données. Par exemple, le texte ou les images utilisés pour remplacer un contenu supprimé doivent éviter la divulgation de données sensibles ou, en l'absence de décision judiciaire, les accusations d'illégalité.

g) Réparation des victimes

Victimes d'activités illégales : La modération de contenu étant axée sur la suppression des contenus, les États devraient en outre considérer les droits des victimes de contenus illégaux, en complément de ces activités. Cela est nécessaire pour garantir un soutien total aux victimes afin d'annuler ou d'atténuer les dommages qui ont été causés.

Victimes de démantèlements injustifiés : Des mesures appropriées sont également nécessaires pour indemniser les victimes d'enlèvements injustifiés et pour éviter que de tels problèmes ne se posent. Les problèmes peuvent être évités en ne retirant pas les contenus si cela peut être évité, en sanctionnant les signalements malveillants et en veillant à ce que les intermédiaires d'internet ne soient pas indûment incités à retirer les contenus. Les États doivent aller au-delà des calculs basés sur les pertes financières. Ils devraient veiller à ce que le coût moral des restrictions excessives de la liberté d'expression, tant pour la victime directe que pour la société dans son ensemble, soit compensé.

7. Traiter les particularités de l'autorégulation et de la corégulation en matière de modération de contenu

L'autorégulation des médias traditionnels implique normalement une entité qui réglemente ses propres décisions éditoriales. L'autorégulation et la corégulation en ligne sont parfois identiques (lorsque les plateformes réglementent leurs propres décisions pour un contenu qui est entièrement sous leur propre contrôle) et parfois entièrement différentes (lorsqu'elles réglementent la communication de leurs utilisateurs, notamment à l'échelle). Par conséquent, les hypothèses basées

[150] Recommandation CM/Rec(2018)2 du Comité des Ministres aux Etats membres sur les rôles et responsabilités des intermédiaires d'internet.

sur l'expérience de l'autorégulation des médias traditionnels sont trompeuses dans le contexte de la plupart des autorégulations des intermédiaires d'internet. Ce fait doit être activement pris en compte dans tout projet d'autorégulation ou de corégulation.

Nous avons vu plus haut que les différentes motivations et structures d'autorégulation et de corégulation ont des impacts significatifs sur la responsabilité et sur l'efficacité des mesures en question. Nous avons également vu que différentes formes d'autorégulation et de corégulation impliquent différentes responsabilités pour l'État. Il est donc crucial, tant pour le respect des obligations en matière de droits de l'homme que pour l'efficacité des mesures mises en œuvre, que le rôle de l'État soit honnêtement reconnu, pour assurer la responsabilité et pour tirer parti de l'expérience de mesures similaires dans ce contexte et dans d'autres.

L'internet nous a donné de nouvelles possibilités fantastiques de parler, d'être entendus et de nous organiser. En effet, il a créé une multitude de nouvelles possibilités d'exercer nos droits humains, notamment nos droits à la liberté d'expression, à la liberté de réunion, à la liberté de pensée et de religion, etc. Toutefois, sans surprise, il crée également des opportunités de diffusion de contenus ou de comportements illégaux ou potentiellement préjudiciables. Les services en ligne (tels que les plateformes de médias sociaux, où les gens publient des messages, des articles, des photos, etc.) ont la possibilité de supprimer, de rétrograder ou de décourager de toute autre manière la diffusion de contenus illégaux ou importuns, ce que l'on appelle la "modération du contenu".

La note d'orientation aborde ces questions de manière générale, à savoir comment mieux comprendre la nature des problèmes spécifiques que la modération de contenu cherche à résoudre, comment garantir une responsabilité appropriée en cas de restrictions des droits de l'homme, que signifient les concepts d'autorégulation et de corégulation, et quelles sont les caractéristiques des outils de transparence qui sont fondamentaux pour garantir que les objectifs sont fixés et atteints.

www.coe.int/**freedom**of**expression**

www.coe.int

Le **Conseil de l'Europe** est la première organisation de défense des droits de l'homme du continent. Il compte 47 États membres, dont tous les membres de l'Union européenne. Tous les États membres du Conseil de l'Europe ont signé la Convention européenne des droits de l'homme, un traité visant à protéger les droits de l'homme, la démocratie et l'État de droit. La Cour européenne des droits de l'homme supervise la mise en œuvre de la Convention dans les États membres.

COUNCIL OF EUROPE

CONSEIL DE L'EUROPE