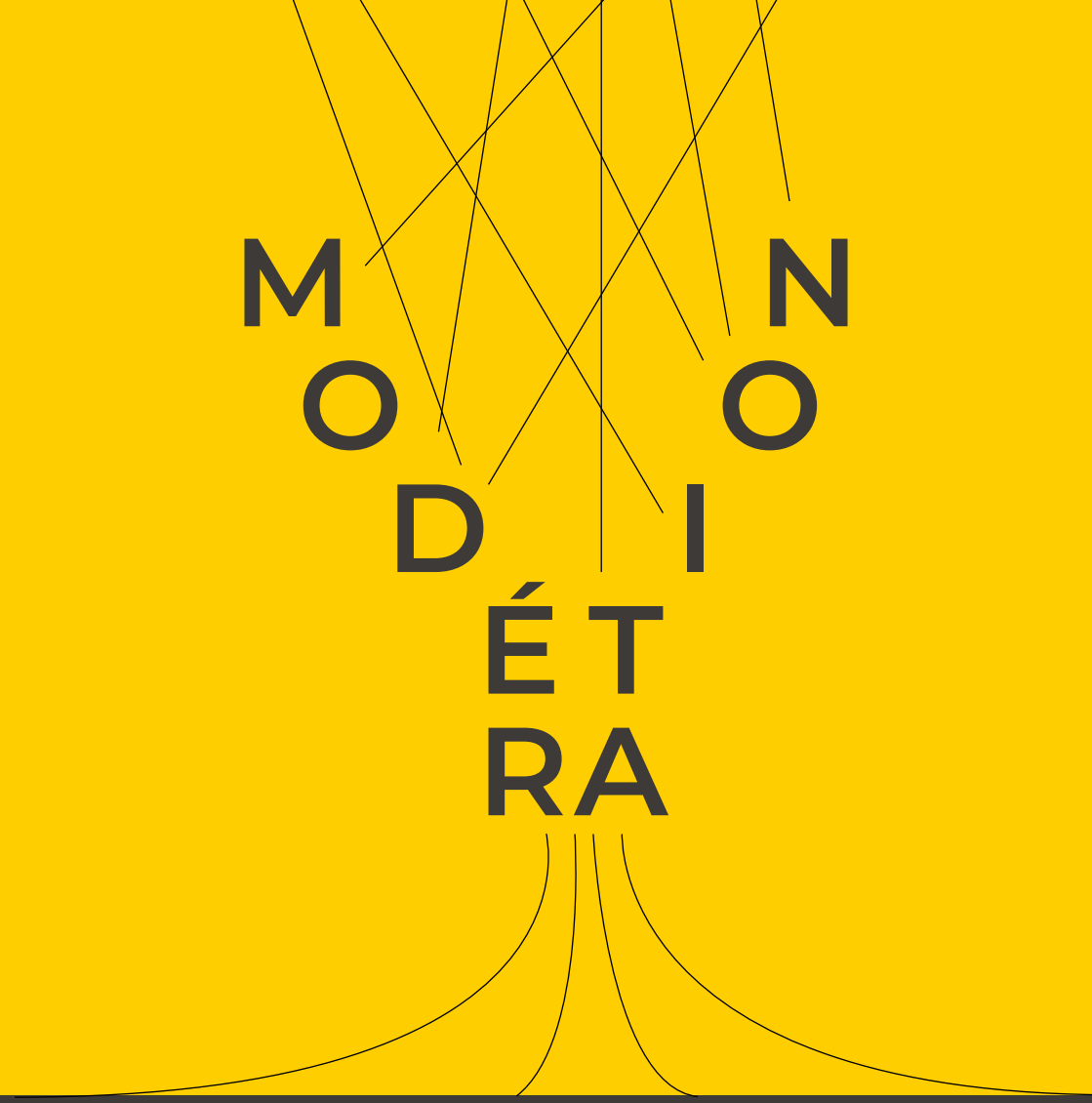


POLITIQUES, INSTITUTIONS ET DÉMOCRATIE
JUIN 2020

MODÉRATION DES CONTENUS : RENOUVELER L'APPROCHE DE LA RÉGULATION

RENAISSANCE numerique

TABLE DES MATIÈRES

CE QU'IL FAUT RETENIR 4

INTRODUCTION
ACCROÎTRE LA PORTÉE
DES POLITIQUES PUBLIQUES 6

PARTIE I
LES DÉFIS DE LA MODÉRATION DANS
UN PAYSAGE EN LIGNE FRAGMENTÉ 15

De la nécessité d'appréhender la diversité des plateformes en ligne 16

Une diversité d'approches de la modération des contenus 17

Les contenus toxiques : un problème pour toutes les plateformes 29

PARTIE II
LES LIMITES DU CADRE LÉGAL ACTUEL 35

La prédominance du modèle de régulation industriel 36

Repenser les indicateurs qui façonnent la régulation 40

PARTIE III
VERS UNE APPROCHE COLLABORATIVE
DE LA MODÉRATION 47

Co-construire des modèles de modération 48

Entretenir une culture de modération en lien avec les utilisateurs 52

CONCLUSION
LA MODÉRATION, UN LEVIER
DE LA DÉMOCRATIE 56

POUR ALLER PLUS LOIN 59

CE QU'IL FAUT RETENIR

- Les politiques publiques ont tendance à ne prendre en compte qu'une poignée de plateformes dans les efforts actuels de régulation des contenus toxiques en ligne. Les discussions politiques en cours ne parviennent pas non plus à saisir les particularités et l'interconnexion de la modération des contenus sur différentes plateformes. Il en résulte que la réponse politique ne s'attaque qu'à une partie du problème.

- La modération des contenus n'est pas seulement une question de suppression. Il s'agit aussi de trouver le bon équilibre, le bon positionnement et le bon processus, en collaboration avec les responsables politiques, la société civile et les utilisateurs finaux eux-mêmes.

- Aujourd'hui, la co-régulation reste un processus bilatéral entre les principaux opérateurs de plateformes et les gouvernements. Si la régulation introduit des responsabilités et des obligations générales calibrées pour les plus grandes firmes mondiales de l'internet, ces mesures auront des effets négatifs disproportionnés sur les autres acteurs, ce qui pourrait par conséquent réduire encore la diversité des plateformes. Pour soutenir un large éventail d'expression en ligne, les cadres de la régulation doivent veiller à ne pas réduire davantage la diversité des plateformes disponibles.

- Une nouvelle approche de la régulation est nécessaire, qui tienne compte des diverses approches de la modération et protège les droits fondamentaux. À cet égard, nous avons besoin d'indicateurs agiles qui nous permettent de mesurer la réactivité des plateformes aux véritables défis de modération auxquels elles sont confrontées, lesquels sont évolutifs.

- La nuance, l'agilité et une large participation des parties prenantes et des utilisateurs finaux sont indispensables à cette nouvelle approche de la régulation. Pour beaucoup, la question centrale est de trouver une façon de dépasser le concept de seuil d'utilisateurs, ou le nombre d'utilisateurs nationaux sur une plateforme. Ce concept est inapproprié, car ce chiffre à lui seul n'illustre pas les défis de modération auxquels une plateforme est confrontée. Renaissance Numérique plaide pour une évaluation des performances de modération des plateformes davantage axée sur les processus.

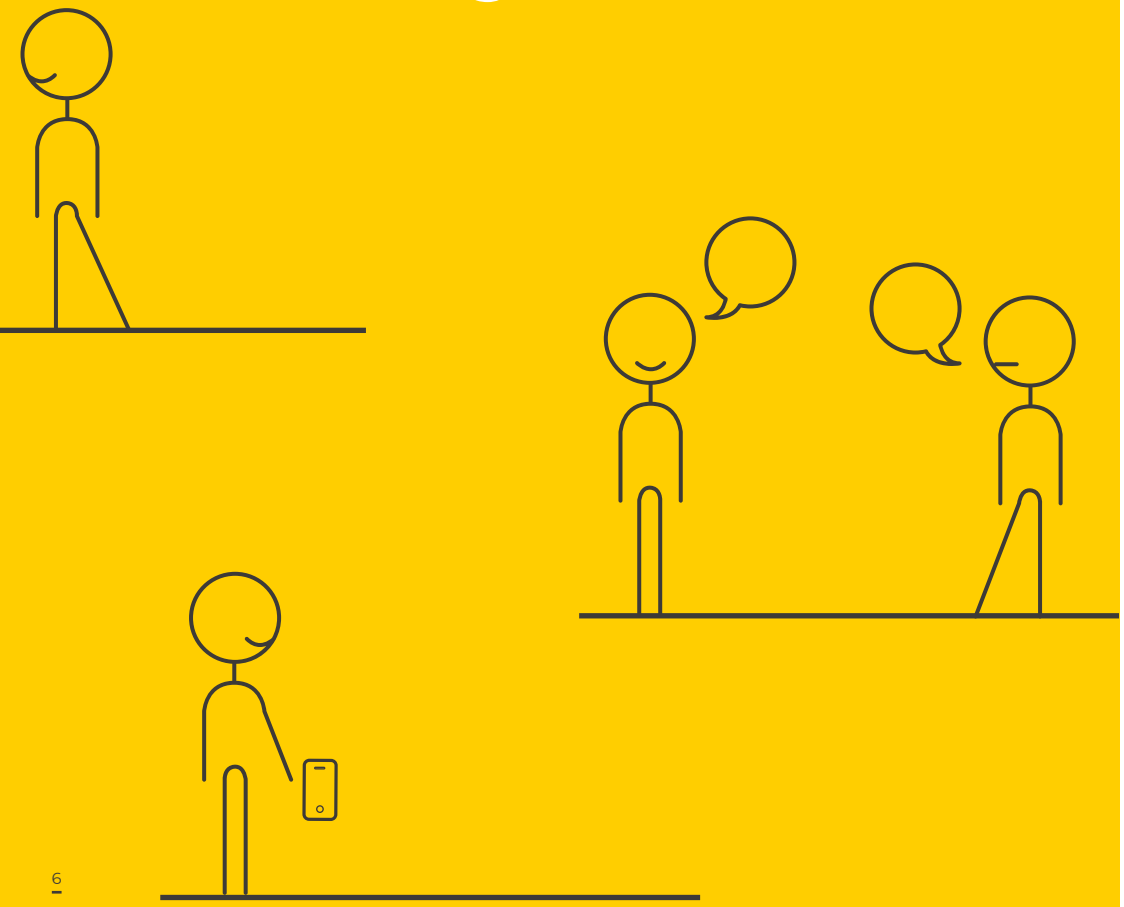
- La notion de co-création de valeur est inhérente aux plateformes qui hébergent des contenus générés par leurs utilisateurs. La contribution substantielle des utilisateurs finaux doit donc se refléter dans la façon dont la modération des contenus des plateformes est gouvernée. Une approche collaborative nécessite de véritables processus discursifs impliquant les utilisateurs finaux, et pas seulement l'externalisation du travail de modération.

- Des structures de gouvernance doivent être mises en place pour faciliter cette participation. Ce type de participation des utilisateurs doit s'inscrire dans un changement de comportement plus large sur les plateformes en ligne, faisant de l'utilisateur final un acteur principal.

- Les autorités publiques doivent renforcer les capacités de toutes les parties prenantes pour permettre une collaboration fonctionnelle et des processus discursifs. Il est de leur responsabilité d'établir un cadre général pour faciliter la collaboration intra et intersectorielle et le partage des connaissances, de travailler avec la société civile, les chercheurs et les experts techniques pour identifier des méthodes efficaces, et de partager ces méthodes avec tous les acteurs et sur toutes les plateformes.

- La future régulation dans ce domaine, en particulier le *Digital Services Act* européen, ne doit pas être façonnée uniquement pour et par les opérateurs de plateformes les plus dominants. La régulation doit viser à aborder la modération des contenus de manière globale, pour tous les services concernés.

INTRODUCTION
ACCROÎTRE LA PORTÉE DES POLITIQUES PUBLIQUES



Les contenus préjudiciables en ligne et la question de savoir comment y remédier existent depuis les origines du web dit « de surface » (ou *surface web* en anglais)[1]. Au fil du temps, les espaces en ligne étant devenus essentiels à nos démocraties, ce problème s'est amplifié au point que les contenus toxiques menacent désormais la libre circulation de l'information et la jouissance de nos droits fondamentaux. Toutefois, à l'heure actuelle, les politiques publiques ont tendance à ne prendre en compte qu'une poignée d'opérateurs de plateformes dans leurs efforts actuels de régulation des contenus toxiques en ligne (qu'il s'agisse de discours haineux, de cyberharcèlement, de désinformation, etc.). Le résultat de cette tendance est que la réponse politique ne s'attaque qu'à une partie du problème. Plutôt que de se pencher sur l'ensemble des opérateurs de plateformes qui hébergent des contenus toxiques, ainsi que sur les importantes interconnexions et les effets de contagion entre les différentes plateformes, le regard des politiques publiques reste fixé sur un groupe d'acteurs défini - notamment Facebook, YouTube et Twitter. En Allemagne, la *Netzwerkdurchsetzungsgesetz* ou « Loi NetzDG », dédiée à la lutte contre les discours de haine sur Internet, s'est clairement orientée vers ces trois acteurs : lors de son élaboration, le ministre allemand de la justice, Heiko Maas, a formé un groupe de travail afin de rencontrer spécifiquement Google, Facebook et Twitter[2]. En France, Laetitia Avia, députée LREM du 8e arrondissement de Paris et porte-parole de la loi française analogue contre les contenus haineux en ligne, a déclaré que cette loi était destinée à s'adresser à « une poignée d'acteurs »[3][4]. Malgré cette préoccupation politique concentrée sur un groupe li-

1 Le web dit « de surface » désigne la zone du World Wide Web qui est accessible au grand public et indexable par les moteurs de recherche.

2 La Loi NetzDG est souvent surnommée « la loi Facebook ».
William Echikson et Olivia Knodt, "*Germany's NetzDG: A key test for combating online hate*", *CEPS Research Report*, novembre 2018. Disponible en ligne : https://bit.ly/2ZnSAeM

3 Conférence « Les réseaux de la haine », 28 janvier 2020, École militaire, Paris.
Les acteurs ne sont pas encore déterminés au moment de l'écriture de cette note (juin 2020). Ces derniers, ainsi que des spécificités complémentaires, devraient être déterminés par décrets dans les prochains mois.

4 Après saisine d'un groupe de sénateurs, le Conseil constitutionnel a jugé la loi Avia substantiellement inconstitutionnelle.
Décision n° 2020-801 DC du 18 juin 2020, *Loi visant à lutter contre les contenus haineux sur Internet*. Disponible en ligne : https://www.conseil-constitutionnel.fr/decision/2020/2020801DC.htm

mité de plateformes, tous les opérateurs hébergeant des contenus générés par les utilisateurs de leurs services sont confrontés au défi des contenus toxiques et doivent être judicieusement pris en compte dans la formulation de la régulation. Une régulation qui ne reconnaît pas la diversité des plateformes et les relations entre elles ne compromet pas seulement son efficacité-même, elle risque également de nuire de manière disproportionnée à certaines plateformes, ainsi qu'à la qualité et à la quantité des espaces disponibles en ligne.

Certes, cette attention politique portée sur une poignée de grandes plateformes s'explique par la nature oligopolistique de ces dernières et leur rôle dans la structuration de l'espace public numérique. Mais des décisions en matière de régulation pourraient avoir un impact disproportionné sur les acteurs les moins à même de répondre à ces nouvelles exigences - ceux dont les moyens sont moindres en termes de ressources humaines, de capacités techniques ou ergonomiques, etc. En outre, certaines nouvelles exigences ne tiennent pas suffisamment compte des particularités des différentes plateformes et ne répondent donc qu'à une partie du problème que la régulation vise à résoudre. Par ailleurs, la modération algorithmique automatisée des contenus toxiques (qui devient *de facto* l'exigence pour les plateformes et qui a connu une forte augmentation ces derniers mois lors de la crise sanitaire du Covid-19[5]) est loin d'être une solution miracle[6]. Ainsi, une approche de la régulation donnée est susceptible de capter involontairement dans ses filets d'autres plateformes que celles initialement visées. Ces défis sont à la fois politiques, techniques et financiers. Si les nouvelles régulations introduisent de nouvelles responsabilités et obligations universelles qui sont calibrées pour les plus grandes plateformes numériques du monde, ces mesures auront des effets négatifs disproportionnés sur les autres acteurs, ce qui pourrait, *in fine*, réduire encore la diversité des plateformes disponibles pour soutenir un large éventail d'expression en ligne. Dans le même temps, ces grands groupes du numérique, en quelque sorte « historiques », restent mieux placés pour s'adapter aux régulations et respecter les obligations de conformité[7]. Afin d'élaborer des méthodes efficaces permettant de résoudre le problème des contenus toxiques dans nos espaces en ligne (et ainsi relever les défis de la modération dans un paysage en ligne profondément fragmenté) il convient de regarder au-delà des principaux opérateurs de plateformes qui façonnent actuellement le débat public. La nuance, l'agilité et une large participation des parties prenantes et des utilisateurs finaux sont indispensables à une nouvelle approche de la régulation.

Les outils et les usages des plateformes numériques sont en constante évolution, de même que les tendances en matière de nature et de propagation des contenus toxiques. Non seulement les propos toxiques sont mouvants dans les espaces en ligne et hors ligne (dialecte, symboles, codes), mais les plateformes utilisées et les activités sur ces dernières se transforment également avec l'émergence de nouveaux outils, fonctionnalités et tendances. La culture du « mème »[8] en est un exemple significatif, qui est pourtant rarement évoqué dans les conversations politiques. Ainsi, les changements sociaux et culturels se superposent aux évolutions techniques, entraînant une mutation rapide des contenus toxiques[9]. Ce n'est là que l'un des aspects du problème. Même ce qui constitue un contenu toxique fait l'objet de débats et de développements continus de la part des plateformes, des responsables politiques, des chercheurs et, bien sûr, des utilisateurs du monde

5 Marc Faddoul, "COVID-19 is triggering a massive experiment in algorithmic content moderation", *Brookings*, 28 avril 2020 : https://www.brookings.edu/techstream/covid-19-is-triggering-a-massive-experiment-in-algorithmic-content-moderation/

6 Hannah Bloch-Wehba, "Automation in Moderation", *Cornell International Law Journal* (à paraître), dernière révision, 29 avril 2020 : https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3521619

7 Nine Principles for Future EU Policymaking on Intermediary Liability, *Center for Democracy and Technology*, avril 2020. Disponible en ligne : https://cdt.org/wp-content/uploads/2019/08/Nine-Principles-for-Future-EU-Policymaking-on-Intermediary-Liability-Aug-2019.pdf

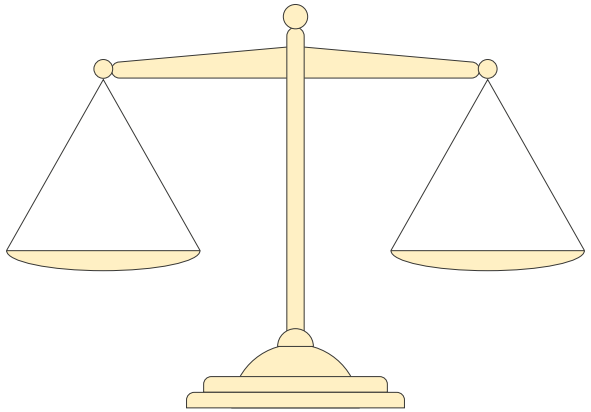
8 Parce qu'ils reposent sur des images adaptées, des symboles et de l'ironie, les « mèmes » sont particulièrement difficiles à modérer. Facebook AI Research a lancé une base de données appelée « *Hateful Memes* », composée de 10 000 mèmes issus de groupes Facebook publics aux États-Unis. "Facebook is using more AI to detect hate speech", *Venture Beat*, 12 mai 2020 : https://venturebeat.com/2020/05/12/facebook-is-using-more-ai-to-detect-hate-speech/

9 Bien entendu, les fonctionnalités avancées et les évolutions des plateformes ne sont pas une condition *sine qua non* de la propagation de contenus toxiques.
"The Hottest Chat App for Teens Is ... Google Docs", *The Atlantic*, 14 mars 2019 : https://www.theatlantic.com/technology/archive/2019/03/hottest-chat-app-teens-google-docs/584857/

entier. Il n'existe pas de définition commune (et certainement pas de définition commune ET opérationnelle) de ce qui constitue un contenu toxique ou nocif sur les plateformes d'hébergement de contenus générés par les utilisateurs. Nous avons par exemple vu les définitions du terme « préjudice » s'élargir tout récemment lors de la pandémie de Covid-19[10]. L'objectif de cette note n'est pas de s'essayer à une définition plus précise de ce qui constitue des contenus toxiques (dont la nature-même peut en fait être poreuse[11]), mais plutôt d'examiner les mécanismes de leur modération et de proposer des pistes d'amélioration. Inévitablement, nous nous concentrons ici sur les contenus interdits par la loi et par les Conditions de service des opérateurs de plateformes (c'est à dire tout matériel constituant une atteinte sexuelle sur mineur, les contenus terroristes, la désinformation, les cybermenaces et le cyberharcèlement, les contenus haineux/diffamatoires/discriminatoires, etc.). Ces contenus toxiques sont en perpétuelle évolution tant dans leurs performances (c'est-à-dire leur façon de se propager en ligne) que dans leurs définitions (la façon dont ils sont perçus), multipliant ainsi les défis pour les régulateurs et pour les opérateurs de plateformes.

En réponse à ces changements, les pratiques de régulation et de modération cherchent à évoluer rapidement, de façon à suivre le rythme du problème. Cependant, toute régulation hâtive comporte ses propres risques[12].

Le futur *Digital Services Act* de la Commission européenne, qui devrait réviser la Directive e-Commerce de 2000, représente l'opportunité de « *renforc[er] nos règles en matière de responsabilité et de sécurité pour les plateformes, les services et les produits numériques* », selon la Présidente de la Commission Ursula von der Leyen[13]. Ce texte vise à poser le cadre d'une gouvernance plus responsable de nos environnements numériques, au centre de laquelle se trouve la question de la modération des contenus. La présente publication vise à nourrir les réflexions de la Commission sur ce sujet.



10 Evelyn Douek, "COVID-19 and Social Media Content Moderation", *Lawfare,* 25 mars 2020: https://www.lawfareblog.com/covid-19-and-social-media-content-moderation

11 Renaissance Numérique (2018), « Fake News, Faire face aux troubles informationnels à l'ère numérique ». Disponible en ligne: https://www.renaissancenumerique.org/system/attach_files/files/000/000/155/original/RN-NOTE_FAKE_NEWS_23mars2018.pdf?1521799239
Renaissance Numérique (2017), « Agir face à la haine sur Internet dans une société collaborative ». Disponible en ligne: https://www.renaissancenumerique.org/system/attach_files/files/000/000/128/original/Note_Seriously_Juillet2017.pdf?1499691042

12 Nous assistons à des évolutions techniques et à l'adoption de nouvelles technologies par la société qui s'effectuent à une vitesse sans précédent. L'adoption d'un grand nombre de plateformes numériques s'est faite en quelques années seulement, alors que celle des technologies précédentes, comme le téléphone ou l'électricité, s'était faite sur des décennies. Cette adoption plus lente avait permis aux régulateurs, à l'industrie et à la société de développer progressivement des comportements et des pratiques de modération.

13 Commission européenne, « Orientations politiques pour la prochaine Commission européenne (2019-2024) »: https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_fr.pdf

LA MODÉRATION EN LIGNE, CE N'EST PAS SEULEMENT DE LA SUPPRESSION DE CONTENU

Les défis de la modération ne se limitent pas à la capacité des opérateurs de plateformes à reconnaître les contenus toxiques. La modération concerne également les décisions prises en matière de conservation et de promotion des contenus, ainsi que d'incitation et de dissuasion de certains comportements. De nombreuses plateformes utilisent des mesures dites «graduelles», par exemple le déclassement et la réduction de la visibilité d'un contenu, la mise en «quarantaine» d'un contenu, le déréférencement d'un contenu, l'ajout d'un label, d'une alerte ou d'informations supplémentaires/qualifiantes, ou même la mise en garde des utilisateurs avant la publication d'un contenu (ce que l'on appelle le «nudging»). La modération peut également consister à fermer des groupes, à suspendre des comptes et à bannir des utilisateurs de la plateforme. Les plateformes innovent considérablement dans ce domaine, et leurs capacités à le faire n'en sont sans doute encore qu'à leurs débuts[14].

En réalité, le simple retrait d'un contenu peut constituer un contournement voire une aggravation du problème créé par ce contenu. La suppression d'un message peut produire une réaction «hydriforme», motivant plusieurs autres messages (de la part d'utilisateurs qui contestent la suppression de leur contenu, ou d'utilisateurs qui sympathisent avec le message original, etc.). Souvent, le contenu qui s'ensuit est encore plus difficile à modérer, dans la mesure où il est rédigé de façon à respecter les conditions d'utilisation juste assez pour rester dans les limites de l'acceptable ou pour contourner la détection. Il peut également arriver qu'un commentaire d'incitation original ne soit pas toxique ou contraire aux conditions d'utilisation, mais que ceux qui s'ensuivent et qui y font référence ou l'évoquent le soient, laissant ainsi aux modérateurs de la plateforme la difficulté de décider s'il faut supprimer le contenu conforme original ou non. La modération par retrait comporte également le risque que la suppression d'un contenu puisse avoir l'effet inverse de celui souhaité : lui donner une légitimité auprès de certaines «communautés d'intérêt» et inspirer sa propagation ailleurs. C'est souvent le cas des théories conspirationnistes, dont la disparition peut servir à renforcer la revendication[15].

En vérité, tout retrait d'un contenu est toujours relatif : il est impossible de retirer entièrement un contenu d'Internet. À la suite des attaques de Christchurch, par exemple, des milliers de personnes ont pu regarder à nouveau la vidéo de la fusillade sans la signaler, la plupart du temps sur des réseaux dits «alternatifs». Et bien que le consortium de partage de *hashes*[16] du GIFCT (Global Internet Forum to Counter Terrorism)[17] soit finalement parvenu à réduire la diffusion des images de la fusillade de Halle, il n'a pas été en mesure les effacer entièrement[18].

Enfin, pour les plateformes et leurs utilisateurs, il peut y avoir des dommages causés à la fois par la suppression et la non-suppression de contenu. Les plateformes attestent subir des réactions négatives de part et d'autre (que ce soit lorsqu'elles retirent du contenu ou lorsqu'elles le laissent en place), y compris de la part d'acteurs de la société civile faisant autorité comme les ONG, qui contestent certaines décisions. Les

14 Twitter teste actuellement de nouveaux paramètres visant à limiter les interactions non souhaitées. Voir : https://blog.twitter.com/en_us/topics/product/2020/testing-new-conversation-settings.html

15 Sam Levin, "'Taking them down fuels it more': why conspiracy theories are unstoppable", *The Guardian*, 28 février 2018 : https://www.theguardian.com/us-news/2018/feb/28/florida-shooting-conspiracy-theories-youtube-takedown

16 Les *hashes* (terme anglais) sont des sortes d'empreintes digitales numériques uniques.

17 Le GIFCT (le forum Internet mondial de lutte contre le terrorisme) a été créé en 2017 par Facebook, Microsoft,Twitter et Youtube, afin de promouvoir le partage d'informations sur les contenus terroristes violents et ainsi faciliter leur suppression sur les plateformes. Le GIFCT a notamment mis en place une base de données pour le partage des *hashes* (empreintes digitales numériques) des contenus terroristes identifiés, afin de faciliter leur retrait. Pinterest, Dropbox, Amazon, LinkedIn et WhatsApp ont depuis rejoint le collectif, entre autres. L'adhésion est également ouverte aux petites plateformes, et les membres de cette initiative s'efforcent de partager leurs ressources (notamment en collaboration avec l'organisation indépendante Tech Against Terrorism). Néanmoins, le GIFCT est souvent critiqué pour son manque de transparence et de contrôle. Voir le site web dédié à cette initiative : https://www.gifct.org/

18 Renaissance Numérique, «Un crime répété, et pourtant : qu'est-ce qui a changé dans notre réponse au terrorisme lié à internet ?», *blog.seriously.ong*, 12 octobre 2019 : http://blog.seriously.ong/un-crime-copie-et-pourtant-quest-ce-qui-a-change-dans-notre-reponse-au-terrorisme-lie-a-internet/

représentants des plateformes expliquent qu'ils ne souhaitent pas être le juge ou le garant de la liberté d'expression[19], les décideurs de ce qui peut apparaître ou non en ligne. Et pourtant, des décisions doivent être prises. Leur travail de modération ne consiste donc pas seulement à supprimer, mais aussi à trouver le bon équilibre, le bon positionnement et le bon processus, avec les responsables politiques, la société civile et les utilisateurs finaux eux-mêmes.

Par-dessus tout, il convient de rappeler que la suppression de contenus constitue un acte politique très fort, en un sens violent, à l'égard de la liberté d'expression, qui constitue un droit fondamental. Le Conseil d'État français l'a rappelé récemment dans le cadre de la loi Avia, en soulignant qu'il s'agit d'un acte « particulièrement radical »[20]. En effet, après qu'un groupe de sénateurs français ait porté l'affaire devant la plus haute juridiction française (le Conseil constitutionnel), le tribunal a estimé que plusieurs articles centraux de la loi, et tout particulièrement le premier article, portaient atteinte à l'exercice de la liberté d'expression et de communication de manière inutile, inappropriée et disproportionnée[21].

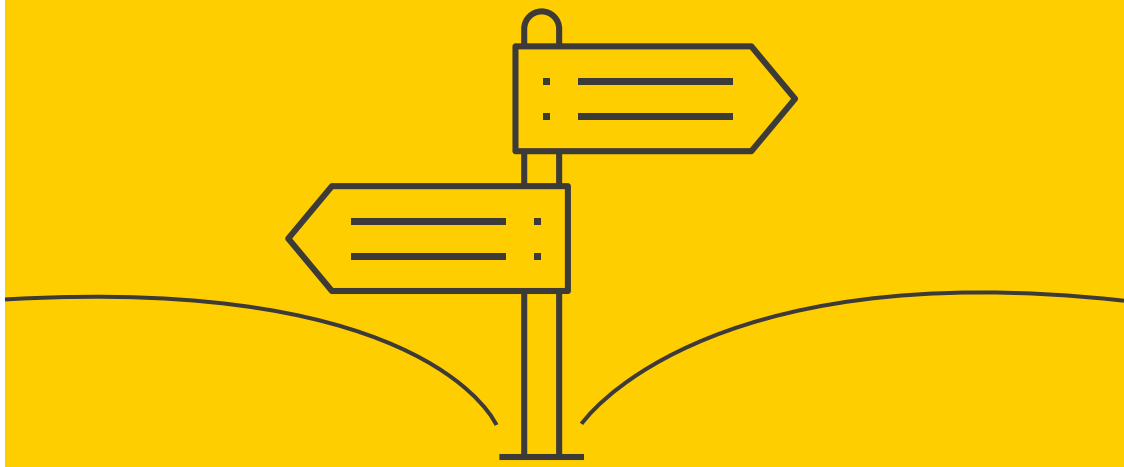
La régulation afférente à la modération des contenus exige la maîtrise d'outils et de pratiques en mouvement constant, la compréhension d'un paysage en ligne divers et fragmenté, et une conscience du fait que la modération des contenus ne peut être ni neutre ni parfaite. Toutefois, une question particulière émerge de ces défis qui se chevauchent, à laquelle cette note cherche à répondre : comment intégrer toutes les plateformes dans le débat sur la modération et faciliter des pratiques de modération qui soient en accord avec nos droits et libertés fondamentaux ?

19 Voir l'entretien complet de CNBC avec le PDG de Facebook, Mark Zuckerberg, du 28 mai 2020 : https://www.cnbc.com/video/2020/05/28/watch-cnbcs-full-interview-with-facebook-ceo-mark-zuckerberg.html

20 Renaissance Numérique (2019), « Lettre ouverte collective appelant à garantir nos libertés publiques dans la proposition de loi visant à lutter contre la haine sur Internet ». Disponible en ligne : https://www.renaissancenumerique.org/system/attach_files/files/000/000/199/original/lettre_ouverte_relative_a_la_proposition_de_loi_visant_a_lutter_contre_la_haine_sur_internet.pdf?1569397597

21 Décision n° 2020-801 DC du 18 juin 2020, *Loi visant à lutter contre les contenus haineux sur internet*. Disponible en ligne : https://www.conseil-constitutionnel.fr/decision/2020/2020801DC.htm

PARTIE I
LES DÉFIS DE LA MODÉRATION DANS UN PAYSAGE EN LIGNE FRAGMENTÉ



DE LA NÉCESSITÉ D'APPRÉHENDER LA DIVERSITÉ DES PLATEFORMES EN LIGNE

Bien que la recherche[22] et les politiques publiques aient tendance à se concentrer sur une poignée d'acteurs seulement, il existe une grande variété de plateformes d'hébergement de contenus générés par les utilisateurs qui sont confrontées au défi des contenus toxiques et qui passent pourtant inaperçues. Nombre de ces plateformes ne sont ni petites ni de niche[23]. Certaines initiatives visent à combler cette lacune. C'est le cas notamment du projet sCAN, mené par un collectif d'organisations de la société civile européenne, qui a examiné certaines plateformes constituant des « refuges » pour les contenus haineux comme RK.com, Gab.ai, RuTube, Telegram, Disqus, Discordance, Spotify, Pinterest et Tumblr.[24]. Malgré cet effort, un travail comparatif plus large est encore nécessaire pour nourrir les politiques publiques. Il existe des différences importantes entre les plateformes : par exemple, le type de contenu hébergé (texte, vidéo, streaming en direct, contenu éphémère), la stratégie de référencement et de classement du contenu (y compris le rôle de l'intelligence artificielle)[25], les services et fonctions offerts (chat privé, marché en ligne, etc.), le modèle commercial de la plateforme (présence de publicités[26] ou non), la taille et la présence géographique de la plateforme, etc. Les différentes méthodes de modération par lesquelles les plateformes abordent le problème des contenus toxiques sont particulièrement pertinentes aux fins de cette analyse.

22 Ce défaut dans la recherche peut être partiellement attribué au défi que représente l'accès aux données. Par exemple, il est relativement facile de récupérer des données sur Twitter. De ce fait, de nombreuses études sont concentrées sur Twitter, ce qui leur confère une pertinence limitée. Le fait que la recherche soit concentrée autour de quelques plateformes et n'ait pas une portée holistique entrave l'amélioration des stratégies de modération et de régulation. La société civile dans son sens le plus large (y compris les chercheurs), en collaboration avec les plateformes, doit mener des recherches approfondies et comparatives entre plateformes.

23 Daniel Carnahan, "For the first time, LinkedIn included data on its moderation efforts in its biannual transparency report", *Business Insider*, 25 novembre 2019 : https://www.businessinsider.fr/us/linkedin-releases-data-on-spam-scams-and-fake-account-removals-2019-11

24 "Beyond the 'Big Three', Alternative platforms for online hate speech", The EU-funded project sCAN– Platforms, Experts, Tools: Specialised Cyber-Activists Network (2018-2020), avril 2019. Disponible en ligne : https://www.voxpol.eu/download/report/Beyond-the-Big-Three-Alternative-platforms-for-online-hate-speech.pdf

UNE DIVERSITÉ D'APPROCHES DE LA MODÉRATION DES CONTENUS

La diversité des plateformes d'hébergement de contenus générés par les utilisateurs doit être prise en compte dans leur régulation. Pour rendre compte de cette diversité, la présente analyse s'appuie sur les travaux de Robyn Caplan dans son rapport intitulé *Content or Context Moderation?*[27]

25 En France, le rapport de la mission conduite par Benoît Loutrel et remis au secrétaire d'État chargé des technologies numériques afin d'éclairer l'approche de la régulation des réseaux sociaux, souligne la nécessité de se concentrer particulièrement sur les « accélérateurs » de contenus, c'est-à-dire les plateformes qui ont une fonction d'ordonnancement des contenus et possèdent donc « *la capacité d'accélérer la diffusion de certains contenus, ou au contraire, de ralentir leur propagation* ». Cette distinction est certes pertinente, mais la présente note ne se limite pas aux accélérateurs de contenu, de nombreuses autres décisions importantes de modération pouvant être prises sans cette capacité.

26 Le secteur de la publicité en ligne peut se retrouver étroitement corrélé à la diffusion de contenus toxiques, souvent en finançant des pages web qui hébergent un tel contenu.
Voir Renaissance Numérique, "Brand safety dans l'écosystème de la publicité programmatique : quel rapport entre les contenus haineux et les marques », *blog.seriously.ong*, décembre 2019 : http://blog.seriously.ong/brand-safety-dans-lecosysteme-de-la-publicite-programmatique-quelle-rapport-entre-les-contenus-haineux-et-les-marques/
Cette tendance, notamment en ce qui concerne la publicité programmatique automatisée, a inspiré un amendement à la loi Avia appelé "Follow the money", qui aurait obligé les annonceurs à rendre publiques au moins une fois par an leurs relations publicitaires. Cette disposition (article 9) a été déclarée non conforme à la Constitution par le Conseil constitutionnel dans sa décision n° 2020-801 DC du 18 juin 2020. Consulter la décision, publié le 25 juin 2020 : https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000042031970&categorieLien=id

27 Caplan, Robyn. "Content or Context Moderation?: Artisanal, Community-Reliant, and Industrial Approaches", *Data & Society*, 2018: https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf

ainsi que sur les témoignages partagés par des représentants de plateformes lors d'un séminaire organisé en février 2020 par Renaissance Numérique[28]. Robyn Caplan développe un cadre théorique distinguant trois typologies : « industrielle », « artisanale » et « communautaire ». Ces catégories sont nécessairement fluides et de nombreux opérateurs de plateformes s'appuient sur des stratégies hybrides, ou bien modifient leurs approches au fil du temps, au fur et à mesure que leurs services et leurs bases d'utilisateurs évoluent.

Le tableau ci-contre est inspiré du travail de Robyn Caplan dans *Content or Context Moderation?* et a été légèrement simplifié et adapté.

L'APPROCHE INDUSTRIELLE : LA VITESSE ET L'ÉCHELLE D'UNE USINE À DÉCISIONS

Dans le cadre d'une modération de type industriel, des dizaines de milliers de salariés appliquent des règles fixées par une équipe dédiée. Les exemples les plus notables sont ceux de Facebook et YouTube. Les équipes s'appuient de plus en plus sur des outils automatisés pour signaler des contenus tels que les discours de haine. Une grande partie des contenus toxiques ou portant atteinte à la vie privée sont ainsi supprimés *ex ante* par des algorithmes ou des outils de filtrage (par exemple, le système d'empreintes digitales numériques utilisé par YouTube, *Content ID*). La quantité de contenus identifiés *ex ante* dépend de la plateforme et du type de contenu. Par exemple, les spams sont souvent identifiés à près de 100 %, et les contenus constituant une atteinte sexuelle à l'encontre des mineurs sont également identifiés par des algorithmes à un pourcentage élevé. Facebook (qui emploie actuellement plus de 30 000 modérateurs) a annoncé avoir identifié automatiquement 80 % des contenus de discours de haine lors

28 Le 14 février 2020, Renaissance Numérique a organisé un séminaire dédié aux défis de la modération des contenus toxiques intitulé « Comment intégrer l'ensemble des opérateurs de plateformes dans le débat sur la modération ? ». L'événement a réuni des représentants de plateformes, des membres de la société civile, des chercheurs et des représentants des institutions publiques françaises. Voir la section « Remerciements » pour la liste complète des participants.

TABLEAU 1 - TYPOLOGIES DES APPROCHES DE LA MODÉRATION

Approche	Les modérateurs	Les utilisateurs	Les outils
Industrielle	• peuvent employer jusqu'à des dizaines de milliers de personnes dans le monde entier ; • de nombreux modérateurs sont des tiers ou des contractuels ; • les équipes de modération sont séparées des équipes de conception et d'affaires publiques ; • le système est normalisé et formalisé (semblable à celui d'une usine).	• les utilisateurs peuvent signaler des contenus toxiques (bien que les modalités dépendent de la plateforme) ; • relations avec la société civile par le biais des programmes "Trusted Flagger", des équipes de fact-checking, etc.	• la grande majorité des contenus est filtrée par des outils automatisés ; • certaines plateformes participent à des collectifs de partage de *hashes* (photoDNA, GIFCT).
Artisanale	• les équipes de modération comptent entre 5 et 200 personnes ; • les modérateurs échangent/se coordonnent avec les autres équipes ; • une place est laissée à la discussion et les décisions sont prises au cas par cas ("manuellement").	• plus de temps est pris par publication et les utilisateurs sont considérés de manière plus globale / selon l'historique de leur activité en ligne ; • le processus de signalement des contenus est similaire à celui des plateformes industrielles.	• recours limité à l'intelligence artificielle, la plupart des contenus sont examinés *ex post* (non filtrés) ; • les plateformes participent éventuellement à des collectifs de partage de hashes, mais y jouent un rôle passif/non stratégique.
Communautaire	• un modèle à plusieurs niveaux, avec une équipe de base composée de quelques salariés, puis différents degrés de participation et de responsabilité volontaires (modèle «en couches d'oignon») ; • quelques politiques transversales, mais les pages/groupes établissent leurs propres règles et les modérateurs respectifs sont chargés de faire respecter ces règles (modèle «fédéral») ; • les modérateurs bénévoles ne sont pas rémunérés.	• tout utilisateur peut devenir modérateur ; • la responsabilité de modérer peut être accrue au fil du temps ; • le signalement par des utilisateurs varie ; souvent, les utilisateurs peuvent se plaindre directement auprès des modérateurs.	• un recours moindre à l'intelligence artificielle, bien que des outils automatisés soient mis à la disposition des utilisateurs et des modérateurs, qui peuvent les utiliser comme bon leur semble.

de leur dernier cycle d'évaluation[29] (contre seulement 38 % en 2018[30], ce qui pourrait mener les experts à tirer la sonnette d'alarme). L'automatisation est une caractéristique fondamentale du modèle industriel. Sans ces capacités technologiques (par exemple dans le modèle artisanal), la modération ne peut se faire qu'*ex post*. Essentiellement, la modération sur ces plateformes est considérée comme industrielle en raison de : leur taille et leur nombre d'utilisateurs, la taille de leurs équipes de modération, leur recours à l'automatisation et à la modération algorithmique *ex ante*, et la séparation entre leurs politiques et la mise en œuvre de ces politiques au sein de l'entreprise. En effet, une caractéristique importante de l'approche industrielle est de séparer les équipes chargées de l'élaboration des politiques de modération de celles chargées de leur application, tant au niveau organisationnel que géographique.

Ces entreprises disposent en général de plus de ressources que les autres, et continuent à développer leurs capacités de modération. Elles ont souvent commencé par des approches artisanales, puis ont expérimenté au fil du temps, en développant des politiques et des systèmes plus formalisés. Robyn Caplan souligne la croissance simultanée des équipes de modération et la nécessité de créer un appareil de prise de décision semblable à celui d'une usine : « *Des concepts complexes comme le harcèlement ou le discours de haine sont rendus opérationnels afin d'avoir une application de ces règles plus cohérente dans l'ensemble de l'entreprise* »[31]. L'approche industrielle comporte, entre autres, la limite suivante : car elle tente de créer une machine à décisions compartimentée, elle n'est pas en mesure de saisir pleinement le contexte entourant les contenus. Ceci conduit à la fois à des « faux positifs » (la suppression de contenus légitimes), et au fait que certains contenus toxiques échappent à la détection.

29 Voir le dernier rapport de transparence de Facebook portant sur les discours de haine : https://transparency.facebook.com/community-standards-enforcement#hate-speech

30 Facebook, "Hate Speech", transparency.facebook.com, 2018 : https://transparency.facebook.com/community-standards-enforcement#hate-speech
Consulté le 31 juillet 2018.

31 Traduit de l'anglais : *"Complex concepts like harassment or hate speech are operationalized to make the application of these rules more consistent across the company".*

L'APPROCHE ARTISANALE : UN EXAMEN AU CAS PAR CAS

Selon la méthode dite artisanale, la modération est normalement effectuée par une équipe de 5 à 200 employés. Les décisions sont souvent prises au cas par cas. Patreon, Change.org, Vimeo, Discordance et Medium sont des exemples de plateformes appliquant ce type de modération. De façon générale, les plateformes artisanales comprennent les grands forums en ligne, les sites web de soutien aux créateurs de contenu, les services de partage de fichiers, etc. Robyn Caplan souligne que ces plateformes sont l'un des principaux moyens d'accès à Internet pour les particuliers du monde entier. Il existe une grande diversité méthodologique au sein même de l'approche artisanale. Toutefois, les équipes de modération artisanale se distinguent non seulement par leur taille réduite, mais également par le fait que la modération soit effectuée en interne par des employés, plutôt que par des services ou des sous-traitants. Les représentants de ces plateformes insistent également sur le recours limité à l'automatisation et aux algorithmes dans la modération des contenus. Ces entreprises et ces organisations (car elles ne sont, en effet, pas toutes des entreprises) pratiquant l'approche artisanale tendent souvent à adopter une approche « manuelle » et approfondie de la modération, et à être plus attentives au contexte dans lequel le contenu est publié. Elles affirment également recevoir moins de signalements de contenus toxiques, ce qui leur permet de consacrer le temps nécessaire à un examen plus minutieux des cas signalés. Bien que certaines comptent des millions d'utilisateurs, il convient de noter que les plateformes artisanales ne sont pas toujours confrontées à la même masse de contenu que les grands acteurs industriels. Enfin, tout en tenant soigneusement compte du contexte, ces plateformes sont limitées dans leur capacité à appliquer les règles de manière cohérente et à grande échelle.

CHANGE.ORG : METTRE L'ACCENT SUR LE DIALOGUE AVEC LES UTILISATEURS[32]

Change.org France, branche de la plateforme de pétitions en ligne Change.org, dispose d'une équipe d'une vingtaine de personnes composée de développeurs, d'ingénieurs, d'un chef de produit et d'une équipe de campagne qui accompagne les auteurs des pétitions. La société possède la certification dite « *B Corp* », une distinction américaine accordée aux entreprises commerciales répondant à des exigences sociétales et environnementales.

Change.org ne s'appuie pas uniquement ou principalement sur l'intelligence artificielle pour la modération des contenus. Selon la Directrice de Change.org France, *« d'abord parce qu'on n'a pas les mêmes ressources que les grosses plateformes, et aussi parce que l'on voit le risque que cela peut comporter en tant que plateforme de pétition dont le but est de favoriser la liberté d'expression»*. Ainsi, leur modération se fait essentiellement *ex post*. Change.org organise de petits groupes d'échanges en interne pour discuter des décisions de modération, et ses équipes essaient de dialoguer avec les créateurs de contenu et avec la communauté d'utilisateurs. Lors du séminaire organisé par Renaissance Numérique le 14 février 2020, la Directrice de Change.org France a souligné que cette pratique du dialogue est d'autant plus essentielle lorsqu'il s'agit de désinformation, étant donné qu'il est difficile de changer d'avis une fois que le contenu a été partagé.

Compte tenu de la taille de la plateforme, cette pratique du dialogue n'est pas facilement modulable et ne peut être pratiquée pour l'ensemble des pétitions. En tant qu'entreprise américaine, la capacité de Change.org à appréhender le contexte français reste limitée : par exemple, lorsque des employés aux États-Unis sont amenés à prendre des décisions concernant du contenu français comportant des subtilités sociales/culturelles.

32 Témoignage la Directrice de Change.org France lors du séminaire organisé le 14 février 2020 par Renaissance Numérique.

PATREON : TENIR COMPTE DE LA PERSONNE DERRIÈRE LE CONTENU

Patreon est une plateforme de financement participatif dotée d'une équipe de six employés à temps plein, au service d'environ 150 000 créateurs dans le monde entier.

Patreon applique un « processus manuel approfondi » de modération qui tient compte du contexte des publications. Par exemple, ses équipes tentent d'examiner l'ensemble du contenu publié par un auteur afin de comprendre l'intention de ce dernier. Selon la plateforme, *« les créateurs sur Patreon dépendent de nous pour leur salaire. C'est une responsabilité énorme que nous prenons très au sérieux. C'est pourquoi toutes les décisions ayant un impact sur le salaire des créateurs sont prises personnellement et au cas par cas. Aucune décision ayant une incidence sur le salaire d'un créateur n'est automatisée. Chaque cas est toujours examiné par un membre de l'équipe "Confiance et sécurité "»*[33].

Le modèle commercial de Patreon, basé sur la logique de l'abonnement, est déterminant dans l'approche de la modération qu'a cette plateforme. Comme l'a exposé l'un de ses représentants : *« la valeur pour une plateforme de chaque nouvel utilisateur est moindre sur une plateforme d'hébergement de contenu gérée par des annonces, par rapport à la valeur de chaque nouveau créateur sur Patreon, qui fonctionne sur la base de revenus tirés des abonnements »*[34]. Toute-

33 Traduit de l'anglais : *"Creators on Patreon depend on us for their paycheck. This is a massive responsibility and one we take very seriously. For this reason, all decisions that impact creators' paychecks are made personally and case-by-case. No decision that impacts a creator's paycheck is automated — each case is always reviewed by a member of the Trust and Safety team."*.
"How Patreon moderates content", *blog.patreon.com*, 25 juillet 2019 : https://blog.patreon.com/how-patreon-moderates-content

34 Colin Sullivan, Chef du service juridique chez Patreon, "Trust Building As A Platform For Creative Businesses", *TechDirt*, 9 février 2018 : https://www.techdirt.com/articles/20180206/11024139169/trust-building-as-platform-creative-businesses.shtml

fois, comme souligné lors d'une interview avec Robyn Caplan[35], les représentants de Patreon émettent encore quelques inquiétudes quant au manque de ressources : *« La raison pour laquelle je pense que la taille est un facteur utile à prendre en compte est qu'elle reflète les ressources dont dispose [une] plateforme pour se conformer à quelque chose »*[36].

L'APPROCHE COMMUNAUTAIRE : UNE MODÉRATION FÉDÉRÉE ET BÉNÉVOLE

L'approche communautaire de la modération est généralement une combinaison de politiques formelles développées par les plateformes et d'actions prises par des modérateurs volontaires. Les plateformes suivant cette approche établissent habituellement un système par lequel un grand nombre de volontaires mettent en œuvre, voire enrichissent, les décisions de politique générale prises par une équipe principale de taille réduite. C'est le cas notamment de Wikipédia, Reddit ou encore Mastodon. Dans ce modèle, la plupart des opérateurs de plateformes comptent sur leurs utilisateurs pour participer au processus de modération. Ce style de modération repose largement sur des bénévoles qui peuvent avoir différents niveaux d'autorité de modération. Selon Robyn Caplan, la caractéristique la plus importante de cette approche est la relation entre l'organisation mère et ses modérateurs bénévoles.

Ces plateformes sont souvent (mais pas toujours) des structures décentralisées, ce qui constitue à la fois des forces et des faiblesses en termes de modération. Un représentant de Reddit a comparé leur modèle à un système fédéral avec des règles à l'échelle du site qui doivent être respectées par les petites sous-communautés[37], mais qui peuvent également être étendues à la discrétion des modérateurs de ces dernières. Les plateformes communautaires donnent la priorité à la prise de décision locale s'attachant au contexte, souvent au détriment de la cohérence et de l'uniformité de l'ensemble de la plateforme. Permettre aux communautés de suivre leurs propres règles permet sans doute une plus grande sensibilité au contexte. Cependant, des désaccords peuvent survenir (et surviennent effectivement) entre les sous-communautés, et les plateformes dotées d'une telle structure fédérale de modération risquent d'apparaître incohérentes et arbitraires dans leur pratique de la modération. Comme l'illustrent les récents événements survenus sur Reddit, la dépendance des communautés peut engendrer des problèmes non négligeables de transparence et de responsabilité : une capture d'écran d'une liste de modérateurs partagée en mai 2020 affirmait que 92 des 500 subreddits[38] les plus importants étaient contrôlés par seulement quatre personnes. David Pierce, rédacteur pour *Protocol*[39], explique que cette liste est trompeuse car les groupes Reddit comptent souvent des dizaines de modérateurs, même s'il concède que la controverse révèle un défaut sous-jacent, à savoir que certaines personnes ont trop de pouvoir sur la plateforme[40]. Le recours au travail bénévole de la communauté des utilisateurs présente d'autres inconvénients majeurs.

35 Caplan, Robyn. "Content or Context Moderation? : Artisanal, Community-Reliant, and Industrial Approaches", Data & Society, 2018 : https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf

36 Traduit de l'anglais : *"The reason why I think size is a useful thing to think about is it's a reflection of the resources available to that platform to actually comply with something."*

37 David Pierce, rédacteur pour Protocol, illustre cette idée de la façon suivante : *« Pratiquement chaque subreddit, une fois qu'il atteint une certaine taille, développe son propre règlement. Il n'y en a pas deux qui soient identiques : vous pouvez avoir un subreddit "Game of Thrones" qui n'autorise pas les mèmes, mais uniquement les discussions sérieuses, et un subreddit concurrent où les mèmes se répandent comme du vin de Dorne. Certains sont impitoyables en matière de formatage et de style, d'autres s'en moquent complètement »*. Traduit de l'anglais : *"Practically every subreddit, once it hits a certain size, develops its own rulebook. No two are alike: You can have a 'Game of Thrones' subreddit that doesn't allow memes, serious discussion only, and a competing one where memes flow like Dornish reds. Some are ruthless about formatting and style, others couldn't care less."*
David Pierce, "Reddit does moderation differently — and it's ignited a war on the platform", *Protocol*, 27 mai 2020 : https://www.protocol.com/reddit-powermods-war?utm_campaign=The%20Interface&utm_medium=email&utm_source=Revue%20newsletter

38 Les 'subreddits' sont les sous-parties du site Reddit, chacune consacrée à un thème spécifique.

39 *Protocol* est un média en ligne développé par l'éditeur du média politique POLITICO.

40 David Pierce, "Reddit does moderation differently — and it's ignited a war on the platform", *Protocol*, 27 mai 2020 : https://www.protocol.com/reddit-powermods-war?utm_campaign=The%20Interface&utm_medium=email&utm_source=Revue%20newsletter

Ces plateformes sont en effet souvent critiquées pour leur dépendance à l'égard de bénévoles non rémunérés, ces derniers pouvant être exposés aux mêmes types de contenus horrifiants[41] que les modérateurs rémunérés employés par les plateformes industrielles. Les modérateurs bénévoles sont également victimes de harcèlement de la part des utilisateurs qui ne sont pas d'accord avec leurs décisions, qui peut aller jusqu'à des menaces de mort[42]. Cela dit, Reddit représente sans doute une extrémité du spectre en matière de modération communautaire, avec une gouvernance de modération globale minimale, voire inexistante. À cet égard, Wikipédia[43] et la plateforme allemande Gutefrage.net[44], font montre d'une plus grande volonté de façonner et de structurer la modération communautaire, par le biais de couches organisationnelles et de politiques prudentes[45].

41 Casey Newton, "Facebook will pay $52 million in settlement with moderators who developed PTSD on the job", *The Verge*, 12 mai 2020 : https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health

42 David Pierce, "Reddit does moderation differently — and it's ignited a war on the platform", *Protocol*, 27 mai 2020 : https://www.protocol.com/reddit-powermods-war?utm_campaign=The%20Interface&utm_medium=email&utm_source=Revue%20newsletter

43 La modération par des bénévoles varie grandement selon les plateformes et les communautés, mais la communauté de Wikipédia est souvent citée comme particulièrement mature à cet égard. Cela a notamment été démontré par la réaction de la plateforme et de ses modérateurs communautaires face à la vague de désinformation lors de la crise sanitaire du Covid-19. Voir :
Omer Benjakob, "On Wikipedia, a fight is raging over coronavirus disinformation", *Wired*, 9 février 2020 : https://www.wired.co.uk/article/wikipedia-coronavirus

44 Gutefrage.net dispose d'un système de modération à plusieurs niveaux, qui vise à protéger les modérateurs bénévoles du travail traumatisant consistant à examiner les contenus les plus toxiques. Les employés de l'entreprise (qui sont eux des modérateurs professionnels, rémunérés) s'appuient sur des algorithmes pour traiter les contenus les plus dangereux, tandis que les modérateurs bénévoles ne s'occupent que du contenu qui passe ce premier filtre et est publié.
Dinar, Christina, 'Digital Streetwork - Wie Online Pädagogisch Agieren ?', 22 novembre 2019 : https://www.belltower.news/debate-dehate-digital-streetwork-wie-online-paedagogisch-agieren-93255/

45 Adi Robertson, "Wikimedia is writing new policies to fight Wikipedia harassment", *The Verge*, 25 mai 2020 : https://www.theverge.com/2020/5/25/21269482/wikimedia-foundation-anti-harassment-code-of-conduct-vote

WIKIPÉDIA : UNE VIRALITÉ À CONTRESENS ?

Wikipédia est financé par la Wikimédia Foundation, une organisation caritative américaine, et est un projet volontaire et collaboratif largement dépendant des dons reçus par la fondation. Le site compte 30 millions de visiteurs uniques et est le quatrième le plus visité en France, et le cinquième dans le monde[46].

Wikipédia repose sur un système hybride d'outils automatisés et d'éditeurs bénévoles[47]. Tous les utilisateurs sont invités à contribuer et à modifier le contenu du site. Le modèle de modération est donc basé sur l'engagement «civique» volontaire des individus. Un représentant de Wikimédia France a fait remarquer que Wikipédia subit de ce fait l'inverse des effets négatifs de la viralité : celle-ci a un effet modérateur, étant donné que plus un article est vu/partagé/modifié, plus il est modéré[48].

Wikipédia s'appuie sur un système de modération à plusieurs couches, qui offre différents niveaux de responsabilité et encourage les participants à s'impliquer progressivement. Pour encourager cette participation, Wikipédia facilite les relations entre ses modérateurs. Des réunions locales hors ligne au sein de plusieurs hubs, ainsi qu'une conférence annuelle, la *WikiCon*[49], sont organisées par la plateforme. Bien entendu, le modèle de Wikipédia n'est pas sans complications. De nombreuses sources évoquent notamment le harcèlement dont souffrent les modérateurs sur la plateforme[50].

46 D'après Alexa.com, consulté le 15 juin 2020 : https://www.alexa.com/topsites/countries/FR

47 Clark, Justin, Robert Faris, Urs Gasser, Adam Holland, Hilary Ross, et Casey Tilton. *Content and Conduct: How English Wikipedia Moderates Harmful Speech*. Berkman Klein Center for Internet & Society, Harvard University, 2019 : http://nrs.harvard.edu/urn-3:HUL.InstRepos:41872342

48 Témoignage du Président de Wikimédia France lors du séminaire organisé le 14 février 2020 par Renaissance Numérique.

49 Témoignage de Christina Dinar, anciennement Project Manager chez Wikimédia Germany. Entretien du 25 février 2020. Pour plus d'informations concernant la *WikiCon* : https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:WikiConvention

50 Adi Robertson, "Wikimedia is writing new policies to fight Wikipedia harassment: Trustees say it hasn't done enough to stop abuse", *The Verge*, 25 mai 2020 : https://www.theverge.com/2020/5/25/21269482/wikimedia-foundation-anti-harassment-code-of-conduct-vote

FRAMAPIAF : LA CONFIANCE DANS L'HUMAIN

Framapiaf est une sous-communauté du site Mastodon modérée par Framasoft, une association d'éducation communautaire à but non lucratif qui fournit des logiciels alternatifs en *open source* ainsi que des outils en ligne. La modération de Framapiaf est assurée par cinq volontaires qui communiquent entre eux et avec les employés de Framasoft selon leurs besoins[51]. Des règles explicites et relativement strictes sont énoncées dans la charte de Framasoft[52]. Les modérateurs peuvent notamment se réserver le droit de prendre une décision sans expliquer pleinement leur raisonnement. Chaque fois que cela est possible (en somme, lorsque les modérateurs jugent l'auteur d'un contenu susceptible d'engager un dialogue positif), ces derniers entament une discussion avec l'auteur sur le contenu et sur les raisons pour lesquelles il ne correspond pas aux normes de leur communauté. Lorsqu'un tel échange n'est pas possible, les modérateurs ont la possibilité de retirer le message et même de bannir l'auteur du groupe.

Framasoft ne repose pas sur une modération algorithmique automatisée. Comme en témoigne une modératrice bénévole de la plateforme : *« Framapiaf est un logiciel libre. De plus, il n'y a pas d'algorithmes, pas d'IA, et il y a peu de couches entre les utilisateurs et les modérateurs. Il n'y a pas de couches intermédiaires. Cette touche humaine donne confiance aux utilisateurs »*[53].

51 Témoignage de Maïtané Maiwann, modératrice bénévole de Framapiaf. Entretien du 31 mars 2020.

52 Voir les Conditions Générales d'Utilisation de Framasoft: https://framasoft.org/fr/cgu/

53 Témoignage de Maïtané Maiwann, modératrice bénévole de Framapiaf. Entretien du 31 mars 2020.

Comme en témoignent ces typologies et ces exemples, il n'existe pas un style de modération unique, loin s'en faut. Il convient à ce titre de noter les grandes variations entre plateformes, et même au sein d'une même plateforme au fil du temps, en ce qui concerne les outils, les stratégies, les relations avec les utilisateurs et les sensibilités politiques. C'est cette diversité que les responsables politiques doivent garder à l'esprit lorsqu'ils élaborent la régulation qui traversera ce paysage.

LES CONTENUS TOXIQUES : UN PROBLÈME POUR TOUTES LES PLATEFORMES

Des tendances et des comportements similaires en matière de diffusion de contenus toxiques peuvent être observés sur les plateformes d'hébergement de contenus générés par les utilisateurs. C'est le cas par exemple de la maîtrise des techniques de codage utilisées pour rester dans les limites de la légalité, ou encore de la multiplication des contenus toxiques à la manière d'une hydre en réponse à la suppression du contenu original. En outre, il existe une porosité importante entre les plateformes. Dans certains cas, un même contenu toxique peut se retrouver sur plusieurs d'entre elles : des liens peuvent être partagés vers des contenus hébergés ailleurs, des fichiers stockés dans des groupes ou des conversations privés, etc. À titre d'exemple, bien que la plateforme Pinterest régule les discours de haine de manière relativement rigoureuse, il a été constaté que des utilisateurs en Italie partageaient par le biais de publications sur Pinterest des hyperliens vers des discours de haine plus explicites stockés ailleurs[54]. De même, les intimidateurs et autres agresseurs peuvent poursuivre leur cible sur diverses

54 *Beyond the "Big Three", Alternative platforms for online hate speech*, The EU-funded project sCAN– Platforms, Experts, Tools: Specialised Cyber-Activists Network (2018-2020), avril 2019 : https://www.voxpol.eu/download/report/Beyond-the-Big-Three-Alternative-platforms-for-online-hate-speech.pdf

plateformes[55]. Afin de concevoir des mesures efficaces pour lutter contre les contenus toxiques, les opérateurs de plateformes et les responsables politiques doivent d'abord reconnaître cette porosité inévitable entre les plateformes.

Certaines plateformes constatent même que leurs services sont intentionnellement abusés ou détournés. Par exemple, des contenus illégaux peuvent être téléchargés et partagés sur un site tiers doté de capacités renforcées en matière de respect de la vie privée, et sur lequel il est impossible d'accéder aux contenus et de les supprimer. Ces plateformes se trouvent en quelque sorte prises entre deux feux. Dans des cas exceptionnels, des contenus viraux illégaux peuvent être sauvegardés sur des plateformes tierces en mode « privé », puis implantés sur des plateformes comme les réseaux sociaux classiques. Comme souligné par un participant au séminaire organisé par Renaissance Numérique le 14 février 2020, ce type de détournement nécessite une coopération accrue entre les plateformes et les autres services en ligne[56]. La littérature montre également qu'à la suite du succès des grandes plateformes « industrielles » dans l'identification et le blocage des comptes soutenant des contenus terroristes et extrémistes, ces contenus se sont déplacés vers d'autres espaces en ligne[57]. Ces « refuges » sont choisis pour leur manque de capacité de modération. Ce fut le cas par exemple de la plateforme JustPaste.it, créée et gérée entièrement par un étudiant polonais[58]. Les utilisateurs individuels et des communautés entières qui sont

55 Renaissance Numérique (2019), « Cyberharcèlement : lecture académique de ce phénomène ». Disponible en ligne : https://www.renaissancenumerique.org/system/attach_files/files/000/000/211/original/_CYBERHARCELEMENT_.pdf?1587545965

56 Témoignage d'un représentant de Dailymotion lors du séminaire organisé le 14 février 2020 par Renaissance Numérique.

57 "Beyond the 'Big Three', Alternative platforms for online hate speech", The EU-funded project sCAN– Platforms, Experts, Tools: Specialised Cyber-Activists Network (2018-2020), avril 2019. Disponible en ligne : https://www.voxpol.eu/download/report/Beyond-the-Big-Three-Alternative-platforms-for-online-hate-speech.pdf

58 "Extremists driven off Facebook and Twitter targeting smaller firms", *The Guardian*, 12 juillet 2017 : https://www.theguardian.com/uk-news/2017/jul/12/extremists-driven-off-facebook-and-twitter-targeting-smaller-firms
"How a Polish student's website became an Isis propaganda tool", *The Guardian*, 15 août 2014 : https://perma.cc/B2GH-5BME

expulsés d'une plateforme peuvent ainsi chercher refuge ou se regrouper sur une autre plateforme. Cette méthode de modération menant à une « dé-plateformisation » est d'ailleurs remise en question, dans la mesure où elle peut avoir pour effet de pousser les auteurs de contenus toxiques vers d'autres plateformes où la modération est plus difficile[59]. Dans le cadre du projet sCAN, il a notamment été constaté qu'*« une migration vers des plateformes comme VK.com ou Gab.ai est souvent annoncée ouvertement sur Facebook et Twitter »*. Ainsi, certaines plateformes se retrouvent à héberger les communautés haineuses s'étant initialement rassemblées et organisées sur les grandes plateformes[60]. Lancé en 2017, le collectif Tech Against Terrorism vise à aider petits et grands opérateurs à protéger leurs services d'une exploitation à des fins terroristes ou extrémistes[61].

La quantité de contenu toxique présent sur les plateformes et la manière précise dont il voyage entre elles demeure une question importante. La recherche de données permettant d'illustrer correctement le phénomène de propagation de contenus toxiques et l'efficacité des efforts de modération des différentes plateformes est perpétuelle. Les rapports de transparence publiés par certaines plateformes visent à fournir ces informations et sont souvent les meilleures références dont disposent les chercheurs et les responsables politiques. À cet égard, il convient de noter que les grandes plateformes « industrielles » ont tendance à mieux respecter les règles en matière de transparence, en grande partie parce qu'elles ont plus de ressources à y consacrer. Néanmoins, de nombreux défis et lacunes dans les rapports

59 N. F. Johnson et al.,"Hidden resilience and adaptive dynamics of the global online hate ecology", *Nature*, 2019 : https://www.nature.com/articles/s41586-019-1494-7
Ryan Greer, "Weighing the Value and Risks of Deplatforming", GNET Insights, 11 mai 2020 : https://gnet-research.org/2020/05/11/weighing-the-value-and-risks-of-deplatforming/

60 Beyond the "Big Three", *Alternative platforms for online hate speech*, The EU-funded project sCAN– Platforms, Experts, Tools: Specialised Cyber-Activists Network (2018-2020), avril 2019 : https://www.voxpol.eu/download/report/Beyond-the-Big-Three-Alternative-platforms-for-online-hate-speech.pdf

61 Renaissance Numérique, « Modération des contenus terroristes : défis techniques, enjeux démocratiques », *blog.seriously.ong*, mars 2020 : http://blog.seriously.ong/moderation-des-contenus-terroristes-defis-techniques-enjeux-democratiques/
Renaissance Numérique, « 3 questions à Jacob Berntsson », *blog.seriously.ong*, mars 2020 : https://blog.seriously.ong/3-questions-a-jacob-berntsson/

de transparence des plateformes rendent l'analyse difficile, en particulier l'analyse comparative ou multiplateforme. Le fait que les différents opérateurs de plateformes aient différentes manières de classer et d'aborder les contenus toxiques, rend difficile la comparaison des données collectées par les différents acteurs. Par exemple, Twitter identifie la « manipulation des plateformes » (l'utilisation de robots), mais ce type de contenu peut être observé et classé différemment selon la plateforme[62]. Cela ne veut pas dire qu'il existe un modèle unique de transparence[63]. Cela signifie simplement que les chercheurs et les acteurs qui sont à la recherche de données comparatives utiles devront surmonter ces divergences. Plus important encore, dans la mesure où la modération ne se limite pas au retrait des contenus, ces rapports de transparence, qui se concentrent souvent sur les retraits, ne fournissent pas une vue d'ensemble complète de la modération. Les données relatives aux retraits et aux demandes de retrait (ainsi qu'aux refus et aux litiges en matière de retrait) constituent la base, mais non la totalité, de la transparence. Des chercheurs ayant examiné les rapports de plateformes en Allemagne à la suite de la loi NetzDG expliquent que ces types de mesures quantitatives auto-déclarées ne sont pas nécessairement un indicateur approprié du succès de la plateforme en matière de traitement des contenus toxiques : *« Le nombre de retraits ou le nombre de plaintes devient une mesure pour évaluer l'efficacité de la loi ; ces retraits sont peut-être inefficaces voire même contre-productifs dans la lutte contre la prédominance globale des discours de haine »*[64].

Le « Code de bonnes pratiques contre la désinformation en ligne », adopté en 2018, et le « Code de conduite visant à combattre les discours de haine illégaux en ligne », adopté en 2016[65], constituent des exemples de collaboration autour de normes d'autorégulation entre la Commission européenne et les opérateurs de plateformes. Ces initiatives ont permis de recueillir des données utiles auprès de leurs signataires initiaux, et visent à mettre en lumière des informations supplémentaires à mesure que d'autres plateformes les rejoignent, que l'application des Codes se renforce et que les définitions et les indicateurs deviennent plus cohérents[66]. Toutefois, il est possible d'améliorer encore et de renforcer ces initiatives. Une évaluation du Code de pratique sur la désinformation publiée en mai dernier a notamment révélé les défis posés par le fait que les 13 plateformes signataires aient différentes façons d'appliquer le Code, ce qui limite son utilité pour les chercheurs. Est également déplorée l'absence d'une compréhension commune et d'une approche harmonisée vis-à-vis du concept de désinformation. Les plateformes non signataires trouvent également les rapports d'auto-évaluation des plateformes signataires non harmonisés et « pas intuitifs » (*"not user-friendly"*). Plus récemment, le Forum multipartite de la Commission européenne ayant contribué à l'élaboration du Code de pratique contre la désinformation en ligne, a demandé à ce que des obligations plus strictes soient imposées aux signataires, citant l'expérience de l'« infodémie » autour

62 Autre exemple : certains contenus sont retirés par les plateformes en réponse à des violations des conditions de service, tandis que d'autres sont retirés parce qu'ils sont illégaux au regard du droit national et ont été signalés par le gouvernement. Certaines plateformes distinguent ces demandes de retrait faites par les gouvernements des demandes de retrait faites par d'autres utilisateurs, tandis que d'autres plateformes ne le font pas.

63 Les principes de Santa Clara sur la transparence et la responsabilité en matière de modération de contenu, rédigés en 2018 par un groupe d'organisations, de défenseurs et d'experts universitaires, définissent les niveaux minimums de transparence et de responsabilité pour les plateformes technologiques en matière de modération de contenus générés par les utilisateurs. Ces principes ne sont pas figés, et la crise de Covid-19 a incité les militants, les experts et les plateformes à réfléchir à leurs évolutions possibles.
"EFF Seeks Public Comment About Expanding and Improving Santa Clara Principles Recommendations Sought from Those Affected by Policies to Moderate", Suppress Speech, Communiqué de presse, 14 avril 2020 : https://www.eff.org/press/releases/eff-seeks-public-comment-about-expanding-and-improving-santa-clara-principles

64 Heidi Tworek et Paddy Leerssen, "An Analysis of Germany's NetzDG Law", A working paper of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, 15 avril 2019 : https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf

65 Le code de conduite visant à combattre les discours de haine illégaux en ligne a été initié avec les grands acteurs que sont Facebook, Microsoft, Twitter et YouTube, qui ont depuis été rejoints par Snapchat, Dailymotion et Jeuxvideo.com.
Voir le site web de la Commission européenne : https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

66 Study for the assessment of the implementation of the Code of Practice on Disinformation, 8 mai 2020 : https://ec.europa.eu/digital-single-market/en/news/study-assessment-implementation-code-practice-disinformation

du Covid-19[67].

Pour le moment, le débat public ne parvient pas à saisir pleinement les particularités et l'interconnexion de la modération des contenus sur les différentes plateformes. Cet échec s'explique en partie par le manque de recherche comparative, lui-même dû aux lacunes et aux limites des rapports de transparence. Nous sortirons certainement de la période actuelle avec de nouvelles idées pour éclairer les décisions politiques (bien que l'étendue des données partagées par les plateformes reste à définir)[68]. À l'avenir, les initiatives ne doivent pas manquer de considérer la modération au sens large, à la fois de façon transversale sur l'ensemble des plateformes, et de manière holistique, c'est-à-dire comme une collection de décisions et de processus (au-delà du simple nombre de contenus supprimés).

67 *« Nous prenons note des annonces et modifications faites par les signataires du code de pratique en matière de lutte contre la désinformation liée au Covid-19 sur leurs réseaux. Cela démontre que, lorsque la volonté est présente, ces acteurs peuvent déployer des solutions à grande échelle pour freiner les contenus préjudiciables sur leurs réseaux. »* Traduit de l'anglais : *"We note the announcements and modifications made by signatories of the code of practice in regard to fighting disinformation related to Covid-19 on their networks. This demonstrates that, where willingness is present, these actors can deploy solutions at scale to curb harmful content on their networks."*
Voir la déclaration conjointe complète, publiée le 15 juin 2020 :
https://m.contexte.com/medias-documents/2020/06/Declaration_medias_desinformation.pdf

68 Voir la lettre conjointe signée par des organisations de la société civile et des chercheurs appelant les plateformes à préserver les données relatives aux décisions prises en matière de contenu pendant la crise sanitaire.
"COVID-19 Content Moderation Research Letter", 22 avril 2020 : https://cdt.org/insights/covid-19-content-moderation-research-letter/

PARTIE II
LES LIMITES DU CADRE LÉGAL ACTUEL

LA PRÉDOMINANCE DU MODÈLE DE RÉGULATION INDUSTRIEL

Les cadres juridiques, de même que les modèles de modération spécifiques aux plateformes, sont souvent construits autour des pratiques des premiers acteurs à avoir atteint un certain pouvoir sur le marché. Ces derniers ont été les premiers à expérimenter et à systématiser des principes de modération, ce qui peut être considéré comme une sorte d'avantage dit du « premier entrant »[69]. En effet, bon nombre de plateformes qui modèrent aujourd'hui les contenus de manière industrielle ont débuté avec des capacités et des stratégies de modération plus artisanales. Au fur et à mesure que leur base d'utilisateurs et que leurs ressources ont augmenté (mais avant l'arrivée d'une forte pression régulatrice autour de la modération des contenus), ces plateformes ont eu l'occasion d'innover et de faire grandir leurs équipes et leurs capacités[70]. Par exemple, Google a pu créer *ContentID*, un outil de détection des violations de droits d'auteur, tandis que Microsoft a établi la norme en matière de suivi de contenus constituant une atteinte sexuelle sur mineur grâce à *photoDNA*[71]. À propos de *ContentID*, James Grimmelmann affirme que ce sont les capacités avancées en matière de calcul et de traitement algorithmique qui ont rendu une telle innovation possible, et qui ont permis des techniques de modération que les législateurs et les régulateurs n'auraient probablement pas pu articuler et imposer eux-mêmes[72].

Ces progrès accomplis par les acteurs dominants sur les technologies de modération automatisées ont fini par établir la norme. Inévitablement, la prééminence de ces capacités de modération a conduit les plateformes innovantes à consolider leurs positions dominantes. En France, nombreux étaient ceux qui craignaient que certaines exigences de la loi Avia n'aient cet effet négatif, notamment l'obligation pour les plateformes de retirer certains contenus en 24 heures, voire en 1 heure (pour les contenus illégaux notifiés par les autorités), sous peine de risquer un an d'emprisonnement et 250 000 euros d'amende. Il s'agit là *de facto* d'une exigence de modération de type industriel. Le rapporteur spécial des Nations unies sur la promotion et la protection du droit à la liberté d'opinion et d'expression, David Kaye, s'est exprimé à ce sujet dans une lettre adressée au gouvernement français en août 2019 : *« Je suis vivement préoccupé par le fait que renforcer le rôle des opérateurs de plateformes en ligne dans la modération de contenu sur Internet pourrait accroître encore davantage "une concentration excessive de la propriété et des pratiques, [ce] qui constituent un abus de position dominante sur le marché" »*[73]. En effet, les plateformes non dominantes et non « industrielles » n'ont généralement pas accès à cette intelligence artificielle qui évolue rapidement, ni à des bataillons de modérateurs humains opérant dans le monde en continu. Ces dernières ne possèdent pas les mêmes ressources que les plateformes industrielles pour leur travail de modération et auraient du mal à respecter les nouvelles exigences en matière de régulation. Des obligations de retrait d'une heure et de 24 heures les obligeraient à mobiliser des équipes 24 heures sur 24 en plus des outils de filtrage algorithmique, et à le faire rapidement. En revanche, comme le souligne David Kaye, les acteurs dominants, eux, possèdent et supervisent en général des technologies de modération devenues désormais essentielles :

69 Voir la définition de David Gotteland, « Comment surpasser l'avantage du premier entrant », Décisions Marketing, No. 21 (Septembre-Décembre 2000), pp. 7-14, *Association Française du Marketing*. Disponible en ligne : https://www.jstor.org/stable/40582911

70 Mike Masnick, fondateur et PDG de Floor64 et rédacteur du blog Techdirt, ne mâche pas ses mots sur le sujet : *« Certains d'entre nous continuent de faire remarquer à l'UE que si ces lois sont conçues pour s'en prendre à Google et à Facebook, elles vont manquer leur cible car elles vont surtout servir à enfermer ces entreprises dans leur position de fournisseurs dominants. C'est parce que ces entreprises sont assez grandes pour gérer la charge réglementaire, alors que les startups et les petits concurrents ne pourront pas le faire et en souffriront »*. Mike Masnick fait ici spécifiquement référence au RGPD et à AdTech.
Foundation for Economic Education, "Google and Facebook Will Just Get Stronger if Regulators Get Their Way, Europe's Experience Shows", 27 août 2019 : https://fee.org/articles/google-and-facebook-will-just-get-stronger-if-regulators-get-their-way-europe-s-experience-shows/

71 Evelyn Douek, "The Rise of Content Cartels: Urging transparency and accountability in industry-wide content removal decisions", *The Knight First Amendment Institute*, 11 février 2020 : https://knightcolumbia.org/content/the-rise-of-content-cartels

72 James Grimmelmann, "The Virtues of Moderation", *Yale J.L. & Tech*, 2015 : https://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1110&context=yjolt

73 David Kaye, « Mandat du Rapporteur spécial sur la promotion et la protection du droit à la liberté d'opinion et d'expression » référénce HCR : OL FRA 6/2019, 20 août 2019 : https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL_FRA_20.08.19.pdf

la technologie d'empreintes digitales numériques comme *ContentID*, ou encore l'accès aux bases de données comme celles du GIFCT. La chercheuse Evelyn Douek insiste sur les dangers de ces bases de données de partage de contenu (qu'elle qualifie de « cartels de contenu »[74]) établies par les plus grandes sociétés du numérique. Dangers en termes de transparence et de responsabilité, de concurrence et même d'efficacité. Certes, il se peut que des catégories comme les spams, les droits d'auteur ou encore les contenus constituant une atteinte sexuelle sur mineur reposent sur des paramètres suffisamment clairs. Toutefois, des catégories comme le discours de haine et l'intimidation risquent de se heurter aux limites de cette méthode. Alors que les responsables politiques et les plateformes continuent à élaborer des catégories de contenus toxiques classifiables et évolutives (catégories supposées suffisamment précises pour être traitées *ex ante* par le biais de bases de données), une attention particulière doit être accordée au contenu de ces bases de données (informations qui ne sont pas disponibles actuellement), ainsi qu'à leur gestion. Il s'agit là de comprendre quelles plateformes en sont propriétaires et ont un rôle décisionnel actif, et quelles plateformes sont les destinataires passifs de cette technologie.

Aujourd'hui, la co-régulation reste un processus bilatéral entre les principaux opérateurs de plateformes et les gouvernements. Assez logiquement, les grands acteurs disposent d'une plus grande capacité de lobbying pour influencer la régulation. De leur côté, les représentants des « autres » plateformes, non dominantes, ne sont pas bien intégrés dans le débat public. Ces plateformes s'appuient généralement sur des équipes d'affaires publique plus petites, si tant est qu'elles en aient. En France par exemple, les petites plateformes se plaignent, tout comme la société civile, de ne pas avoir été suffisamment consultées dans le processus législatif dans le cadre de la loi contre la manipulation de l'information de 2018, et plus récemment dans le cadre de la loi Avia sur la cyberhaine[75]. Pour illustrer à quel point le lobbying

74 Evelyn Douek, "The Rise of Content Cartels: Urging transparency and accountability in industry-wide content removal decisions", *The Knight First Amendment Institute,* 11 février 2020 : https://knightcolumbia.org/content/the-rise-of-content-cartels

75 Témoignages de participants lors du séminaire organisé le 14 février 2020 par Renaissance Numérique.

peut être critique dans la régulation des contenus, il est intéressant de noter qu'en Allemagne, le lobby de l'industrie du jeu est parvenu à faire supprimer certaines des obligations du projet initial de la loi NetzDG[76].

La prépondérance du modèle de modération industriel engendre également des conséquences sociales et politiques. Les cadres de modération construits autour de ces modalités industrielles tendent en effet à favoriser une modération *ex ante*, ainsi qu'une approche relativement conservatrice des contenus se trouvant à la limite de l'acceptable. Ceci augmente les risques de « faux positifs », c'est-à-dire que des contenus légitimes soient supprimés. Malgré les appels lancés par les plateformes, la société civile et même les États contre une « surveillance générale »[77], les récentes tendances en matière de régulation semblent pousser les plateformes vers ce type d'approche, dans la mesure où des délais, des obligations et des sanctions stricts (comme le retrait des contenus terroristes en une heure)[78] obligent effectivement à un filtrage automatisé des contenus[79]. Dans sa décision du 18 juin 2020, le Conseil constitutionnel français a évoqué le risque que la loi Avia n'encourage les plateformes à retirer des contenus légitimes, portant ainsi atteinte de manière excessive à la liberté d'expression. Le juge constitutionnel a ainsi estimé qu'en l'absence de motifs précis d'exonération

76 Hartleb, Florian. "Lone Wolves: The New Terrorism of Right-Wing Single Actors", *Springer*, 2020

77 *"General monitoring"* en anglais.
D9+ Non-Paper on the creation of a moderation regulatory framework for the provision of online services in the EU: https://www.gov.pl/web/digitalization/one-voice-of-d9-group-on-new-regulations-concerning-provision-of-digital-services-in-the-eu

78 La loi française sur la cyberhaine votée en mai 2020 est la dernière en date ayant tenté d'imposer un délai d'une heure pour le retrait de « contenus terroristes ». Accéder au texte en ligne : http://www.assemblee-nationale.fr/dyn/15/dossiers/lutte_contre_haine_internet

79 Hannah Bloch-Wehba, Automation in Moderation, Cornell International Law Journal (à paraître), dernière révision, 29 avril 2020 : https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3521619
Renaissance Numérique s'est joint à d'autres membres de la société civile française dans une lettre ouverte adressée aux législateurs français, « Lettre ouverte relative à la proposition de loi visant à lutter contre la haine sur Internet », juillet 2019 : https://www.renaissance-numerique.org/publications/lettre-ouverte-relative-a-la-proposition-de-loi-visant-a-lutter-contre-la-haine-sur-internet

de responsabilité, les sanctions prévues en cas d'inaction sur les contenus signalés dans le délai prescrit inciteraient les plateformes à retirer les contenus qui leur sont signalés *« qu'ils soient ou non manifestement illicites »*[80].

Le cadre juridique actuel, construit autour de l'approche industrielle de la modération, renforce la prédominance des acteurs déjà dominants (ceux qui disposent de capacités techniques, de ressources humaines et de pouvoir de lobbying). Il incite également l'ensemble des plateformes à avoir recours à des pratiques de modération *ex ante* de plus en plus automatisées, qui risquent de mener à une censure des contenus légitimes. Une nouvelle approche de la régulation est nécessaire et il est impératif que celle-ci tienne compte de la diversité des approches de la modération et protège les droits fondamentaux, notamment en évitant le piège d'une surveillance générale et d'une restriction excessive de la liberté d'expression.

REPENSER LES INDICATEURS QUI FAÇONNENT LA RÉGULATION

Afin de commencer à réguler la modération des contenus, des indicateurs doivent être établis afin d'identifier à la fois les acteurs qui relèvent de la régulation, ainsi que les activités et les aspects qui doivent être surveillés. Les efforts de régulation actuels visant à s'emparer des plateformes industrielles finissent par attirer d'autres acteurs dans leurs filets, tout en ne saisissant pas la complexité du problème. Il est donc nécessaire de s'interroger sur les indicateurs qui alimentent actuellement la politique publique qui, à son tour, détermine le cadre juridique et les modalités de modération de ces plateformes.

À cet égard, il est important de noter que les indicateurs actuels varient selon les pays et les actes législatifs. Il est donc nécessaire d'harmoniser ces indicateurs de même que les nouvelles approches de la régulation au niveau européen. Bien que l'Allemagne et la France aient tenté d'adopter leurs propres législations avant l'arrivée de la loi européenne sur les services numériques (le *Digital Services Act*), une telle approche, qui impose des politiques de modération disparates dans les différents États membres, est moins efficace et fait moins autorité. Un minimum d'harmonisation est essentiel au traitement de contenus toxiques. Une compréhension et une approche communes du problème pourraient être obtenues en partie au travers d'un cadre reposant sur des indicateurs axés sur les processus.

Une question centrale pour beaucoup est de savoir comment dépasser le concept de seuil d'utilisateurs, soit le nombre d'utilisateurs dans un pays donné sur une plateforme au-delà duquel le service sera soumis à une régulation (en France, le nombre généralement évoqué est de 5 millions). Ce concept est inadapté, ce chiffre à lui seul n'illustrant pas les défis de modération auxquels les plateformes sont confrontées. Des critères qualitatifs plutôt que quantitatifs sont indispensables. Par exemple, la plateforme met-elle en place des mesures préventives ? Son modèle économique encourage-t-il la viralité ? L'expérience utilisateur s'appuie-t-elle sur des « zones d'ombre » pour tromper intentionnellement les utilisateurs ? La question du seuil d'utilisateurs a été débattue et finalement adoptée au sein de la loi NetzDG allemande (le seuil retenu étant celui de 2 millions d'utilisateurs inscrits en Allemagne). Il est toutefois intéressant de noter que l'« article 230 », la mesure législative analogue aux États-Unis, ne comporte pas de tel seuil. Si le seuil d'utilisateurs n'est pas un concept idéal ou universel, peut-être faut-il réfléchir plutôt aux modèles économiques des plateformes, à la façon dont elles sont conçues, ou à d'autres caractéristiques spécifiques. La question des indicateurs appropriés et efficaces n'est pas simple dans cet espace, où la régulation doit tenir compte de la grande variété d'opérateurs

80 Citation complète : *« Compte tenu des difficultés d'appréciation du caractère manifestement illicite des contenus signalés dans le délai imparti, de la peine encourue dès le premier manquement et de l'absence de cause spécifique d'exonération de responsabilité, les dispositions contestées ne peuvent qu'inciter les opérateurs de plateformes en ligne à retirer les contenus qui leur sont signalés, qu'ils soient ou non manifestement illicites. Elles portent donc une atteinte à l'exercice de la liberté d'expression et de communication qui n'est pas nécessaire, adaptée et proportionnée. Dès lors, sans qu'il soit d'examiner les autres griefs, le paragraphe II de l'article 1er est contraire à la Constitution. »*
Décision n° 2020-801 DC du 18 juin 2020, *Loi visant à lutter contre les contenus haineux sur internet*. Disponible en ligne : https://www.conseil-constitutionnel.fr/decision/2020/2020801DC.htm

de plateformes confrontés à des défis de modération très différents.

Il est donc impératif de développer des indicateurs agiles qui nous permettent de mesurer la réactivité des plateformes aux problèmes réels auxquels elles sont confrontées, lesquels sont évolutifs. Une régulation qui peut s'avérer utile à un moment donné par rapport à une question spécifique, peut ne pas être en mesure de résoudre des défis futurs. La « Mission Facebook » menée en 2019 en France s'est concentrée sur le classement algorithmique de contenu, ce qui est compréhensible si l'on pense aux problèmes de modération rencontrés par Facebook et YouTube, qui sont souvent liés à la curation et à la viralité des contenus toxiques. Toutefois, d'autres plateformes ne reposent pas sur la curation algorithmique et font tout de même face à un certain niveau de toxicité et de viralité. Ainsi, une régulation établie autour d'une seule caractéristique technique comme le classement de contenu ou la diffusion en direct (des attributs qui ne sont pas nécessairement ou exclusivement la source *per se* de la toxicité des contenus) aura toujours des limites.

L'encadré ci-contre propose des indicateurs pour évaluer la performance des plateformes en termes de modération de contenu. Ces indicateurs cherchent à dépasser le concept de seuil d'utilisateurs et à remettre en question les divers processus et pratiques de modération. Il convient toutefois de noter que toutes les améliorations dans le domaine de la modération ne passeront pas par la régulation. Comme cette analyse tend à le montrer, une régulation trop spécifique risquerait de causer un préjudice involontaire à certaines plateformes en raison de leurs particularités, de la diversité des services qu'elles proposent, et de l'évolution rapide de leurs caractéristiques et des usages qui en sont fait. Les indicateurs proposés ici, classés en cinq catégories, mettent l'accent sur des principes généraux plutôt que sur des méthodes explicites. Les régulateurs pourraient utiliser ces indicateurs pour mesurer l'engagement des plateformes en matière de modération des contenus, sans pour autant imposer de méthodes restrictives concernant la manière dont les opérateurs de plateformes devraient concrétiser cet engagement. Différentes méthodes sont proposées sous ces principes généraux, suggérant des politiques que les plateformes pourraient choisir de mettre en œuvre pour adhérer auxdits principes. Bien entendu, ces suggestions ne doivent pas être considérées comme définitives, mais comme un point de départ en vue d'une collaboration. Ce processus n'en est qu'à ses débuts et le tableau ci-dessous cherche à élargir la discussion plutôt qu'à la restreindre.

DES INDICATEURS PERMETTANT D'ÉVALUER LA PERFORMANCE DES PLATEFORMES EN MATIÈRE DE MODÉRATION DE CONTENUS EN EUROPE

TRANSPARENCE ET RESPONSABILITÉ

- Mise en œuvre de principes et de processus de modération transparents et explicites.
- Publication de rapports de transparence clairs, complets et réguliers.
- Examen de cas individuels permettant d'exposer la logique et le bien-fondé des décisions (examen à la fois qualitatif et quantitatif).

Les plateformes pourraient :

- Fournir des conditions de service claires et accessibles dans toutes les langues dans lesquelles leur service est offert.
- Partager des données pertinentes/appropriées avec les chercheurs, y compris les algorithmes principaux et autres décisions qui influencent les flux de contenu.
- Permettre l'accès à des données agrégées « brutes » à des fins d'analyse (les seuls chiffres agrégés étant difficiles à vérifier).
- Ouvrir leurs API[81] directement aux chercheurs et aux régulateurs.

INVESTISSEMENT GLOBAL DANS LA MODÉRATION

- Maximisation du montant des ressources investies dans la modération en fonction des capacités de l'opérateur.

81 De l'anglais « *application programing interface* », des interfaces de programmation d'applications.

- Investissement dans la modération humaine et dans les humains qui la sous-tendent (sécurité, rémunération).
- Investissement dans le partage des connaissances entre les équipes.

Les plateformes pourraient :

- Accroître leur investissement dans la modération humaine, en particulier dans les modérateurs humains locaux et dans leur formation et leur bien-être continus, en offrant de bonnes conditions de travail et un soutien approprié.
- S'efforcer d'impliquer dans la modération des personnes d'origines et d'expériences de vie diverses en tant que gestionnaires de communauté, experts/conseillers, bêta-testeurs pour les décisions de conception, etc. Cela pourrait impliquer un investissement dans la rémunération d'acteurs externes.
- Investir dans le partage des connaissances au sein de l'organisation, par exemple en mettant en relation les concepteurs et les équipes d'affaires publiques avec les équipes de sécurité, etc.

DIALOGUE AVEC LES UTILISATEURS

- Mise en place de systèmes de consultation mutuelle avec les utilisateurs.
- Mise en œuvre de systèmes de recours et de réparations clairs, transparents et rapides.

Les plateformes pourraient :

- Encourager les utilisateurs à signaler les contenus toxiques.
- Encourager la participation des utilisateurs à la modération. Par exemple, par le biais de systèmes d'engagement progressif.
- Fournir un soutien humain aux utilisateurs qui déposent des plaintes.
- Mettre en place des mécanismes d'appel et de recours transparents, efficaces et rapides, facilement compréhensibles et accessibles (en 3 clics maximum)[82]. Parmi les mesures quantitatives de la réactivité aux plaintes, nous pouvons citer :
  - — la vitesse à laquelle une décision a été prise concernant un contenu signalé (y compris la décision de désactiver l'accès au contenu et pas nécessairement la décision finale) ;
  - — la rapidité avec laquelle les autorités compétentes sont informées en présence d'un contenu manifestement illégal.
- Informer les utilisateurs lorsqu'une décision de modération est prise concernant leur contenu et inclure des informations adéquates sur ce qui a mené à la décision, la règle spécifique qui a été enfreinte, la manière dont les lignes directrices sur la modération des contenus ont été interprétées, les mesures qui seront prises et des instructions claires sur la manière d'introduire un recours.
- Fournir des ressources pédagogiques afin d'aider les utilisateurs à comprendre la logique qui sous-tend les décisions prises.

ATTENTION À L'ÉGARD DE L'EXPERTISE ET DU CONTEXTE LOCAUX

- Inclusion de la société civile et d'experts du contexte local et disposant d'une expertise pertinente.
- Partage des données et des connaissances avec les chercheurs et les autres parties prenantes.
- Engagement à améliorer la logique décisionnelle et la capacité des chercheurs à accéder aux données.

Les plateformes pourraient :

- Solliciter l'expertise de la société civile dans les diverses étapes du processus de conception et de modération (lors de la création de normes communautaires, lors de la conception de produits et d'expériences utilisateurs, lors de la prise de décisions

82 En Allemagne, des outils de signalement « enterrés » ont été jugés comme une entrave au signalement par les utilisateurs sur Facebook. Heidi Tworek et Paddy Leerssen, "An Analysis of Germany's NetzDG Law", A working paper of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, 15 avril 2019 : https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf

de modération, etc.). Pour avoir du sens, toute consultation doit aller au-delà du simple fait de s'appuyer sur les organisations de la société civile pour signaler les contenus (comme cela est le cas par le biais des programmes « *Trusted Flagger* ») et sur les experts pour vérifier les faits. Dans le même temps, ces relations avec la société civile ne sauraient être imposées aux plateformes dont les ressources sont limitées.

- Consulter des experts du contexte local afin de prendre des décisions qui soient culturellement éclairées et qui évitent le problème de la prise de décision « extraterritoriale ».

ALIGNEMENT SUR LES DROITS ET LIBERTÉS FONDAMENTAUX

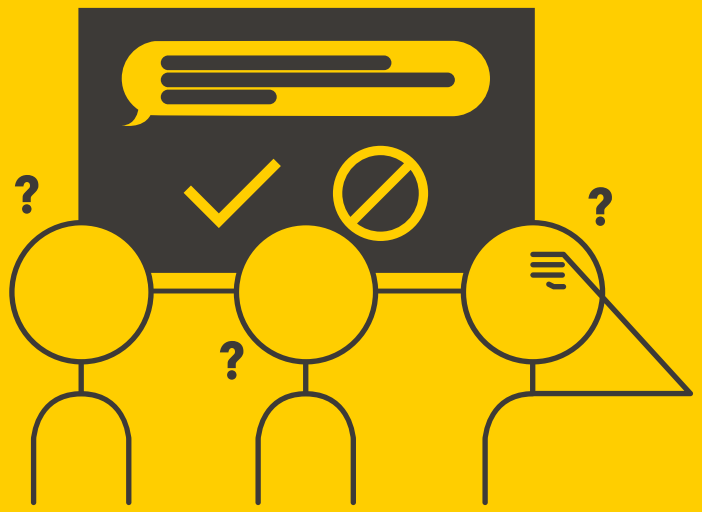
- Mise en place de mécanismes de recours et de réparations qui respectent les droits de la défense.
- Récupération des contenus retirés par erreur.
- Traitement approprié des données des utilisateurs conformément au GDPR et aux cadres juridiques applicables.

Les plateformes pourraient :

- Investir dans le développement de technologies et de matériels de communication destinés à la sécurité des utilisateurs.
- Investir dans le développement de technologies et de matériels de communication destinés à l'éducation et à la résilience des utilisateurs.
- Mettre en place des mesures de conception appropriées pour différents publics (par exemple pour les enfants, les journalistes, etc.), et développer ces ressources et technologies en collaboration avec les utilisateurs eux-mêmes.

Bien qu'ils soient loin de constituer une approche définitive, ces indicateurs ont vocation à ce que les pratiques en termes d'évaluation de contenus dépassent le concept de seuil d'utilisateurs et de nombre de démantèlements, et à ce que l'évaluation de la performance des plateformes en termes de modération s'oriente davantage vers les processus.

PARTIE III
VERS UNE APPROCHE COLLABORATIVE DE LA MODÉRATION



CO-CONSTRUIRE DES MODÈLES DE MODÉRATION

Un défi majeur en matière de modération des contenus est l'application de systèmes et de cadres à grande échelle, et pour des utilisateurs divers à travers de grands territoires culturels/juridiques/sociaux. À cet égard, la contextualisation est essentielle pour comprendre la signification des contenus et la façon dont ils sont perçus. La modération doit tenir compte du contexte local et s'appuyer sur une expertise appropriée. Cette stratégie n'est spécifique à aucune des trois typologies de modération, mais nécessite une approche collaborative. La collaboration se manifestera nécessairement différemment selon la plateforme. Elle peut consister, par exemple, à impliquer des experts locaux, à faire participer la société civile, à travailler avec des journalistes et des *fact-checkers*, ou à partager des ressources pédagogiques ou de renforcement de la résilience qui sont pertinentes dans le contexte local.

Lors du séminaire organisé par Renaissance Numérique le 14 février 2020, plusieurs participants ont exprimé leur souhait de voir se constituer un groupe de discussion entre les plateformes et les organisations de la société civile en France pour partager leurs connaissances et débattre des défis de la modération dans le contexte national[83]. Renforcer la capacité à prendre des décisions de modération au niveau local permettrait d'éviter les failles de modération qui découlent de décisions « extraterritoriales » ou de reléguer une décision au siège de la plateforme (souvent aux États-Unis) où les personnes qui prennent les décisions ne comprennent pas nécessairement tous les éléments contextuels nécessaires à une décision éclairée. Les opérateurs de plateformes ne veulent pas jouer le rôle de juge, en particulier dans les domaines pour lesquels ils manquent d'expertise, et sur des questions culturelles sensibles comme les vêtements religieux, les questions LGBTQ, etc. Ces questions sont en effet difficiles à traiter. En France, certains craignaient que les approches brutales prévues par la controversée loi Avia soient ineptes. Alors que l'association Inter-LGBT dénonce les « raids » ou autres attaques en ligne à l'encontre des contenus LGBTQ (comportements que la loi visait à combattre avec force), cette dernière craignait que les contenus LGBTQ soient retirés de manière injuste et sans explication suffisante ou sans recours approprié en vertu de la nouvelle loi. L'association a également fait part de ses inquiétudes quant au fait que la loi obligerait les mineurs à se manifester s'ils souhaitent déposer une plainte conformément à la procédure légale qui était prévue par la loi. Véronique Godet, co-présidente de SOS Homophobie, s'était à l'époque engagée à rester vigilante sur les modalités d'application du texte : *« quelle assurance avons-nous aujourd'hui que les contenus supprimés sont bien haineux ? Pour l'instant aucune »*[84].

Un organe séparé rassemblant les plateformes et la société civile serait également vraisemblablement bénéfique aux observateurs et aux chercheurs de la société civile, leur donnant accès aux données et aux mécanismes de prise de décision interne. L'*Oversight Board* de Facebook, non sans controverse, semble répondre à certaines de ces questions. TikTok et Twitch prévoient de suivre ce modèle[85] et il est probable que d'autres plateformes feront de même. Ces « conseils de médias sociaux »[86], lorsqu'ils seront effectivement mis en œuvre, pourraient offrir des avantages aux plateformes et aux utilisateurs en remplaçant la prise de décision *ad hoc* par un système plus transparent, responsable et conforme à la loi. Toutefois, ces structures ne devraient pas remplacer les prérogatives de la justice dans un État de

83 La création d'un conseil de surveillance indépendant par des acteurs de la société civile a été évoquée en Allemagne dans le cadre de la loi NetzDG mais n'a pas été concrétisée pour le moment. Témoignage de Christina Dinar, ancienne Directrice de projet chez Wikimédia Germany. Entretien du 25 février 2020.

84 Hervé, Elodie. « Les associations LGBT inquiètes après le vote de la loi Avia contre la haine en ligne », *Têtu*, 13 mai 2020 : https://tetu.com/2020/05/13/les-associations-lgbt-inquietes-apres-le-vote-de-la-loi-avia-contre-la-haine-en-ligne

85 Voir TikTok Newsroom, 18 mars 2020 : https://newsroom.tiktok.com/en-us/introducing-the-tiktok-content-advisory-council

86 "Social Media Councils, from Concept to Reality", Stanford Digital Policy Incubator Conference Report, 1-2 février 2020 : https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/gdpiart_19_smc_conference_report_wip_2019-05-12_final_1.pdf

droit[87]. En outre, ces conseils ne sont pas un outil suffisant pour obtenir une participation massive des utilisateurs : ce ne sont pas des mécanismes de collaboration et de discours ascendants avec les utilisateurs. La société civile et les utilisateurs au sens large devraient pouvoir exprimer leurs préoccupations et contribuer aux politiques de modération des plateformes. La notion de co-création de valeur est inhérente aux plateformes d'hébergement de contenus générés par les utilisateurs : en raison du rôle important qu'ils jouent dans le débat démocratique et de la mesure dans laquelle les utilisateurs finaux contribuent à leur construction[88], ces espaces ne peuvent être simplement considérés comme les territoires d'entreprises privées. Cette contribution substantielle doit se refléter dans leur gouvernance, notamment en ce qui concerne la modération des contenus. L'approche descendante d'un « conseil de surveillance », d'une part, et l'approche plus démocratique et communautaire préconisée dans cette analyse, d'autre part, ne sont pas mutuellement exclusives. De fait, elles peuvent même être complémentaires. Il convient de noter que les conseils de médias sociaux risquent de renforcer l'importance des plus grands acteurs, soit les premiers à s'être lancés dans ce nouvel espace technopolitique : Facebook, TikTok et Twitch, développé par Amazon. Ils peuvent également avoir des conséquences sur la liberté d'expression, en homogénéisant la manière dont les contenus sont régis, comme le souligne Kate Klonic[89]. Renforcer la communication entre les plateformes et les utilisateurs, et impliquer largement les utilisateurs dans le processus de modération, constitue une façon d'atténuer ces risques.

Tout ceci nécessite de construire au préalable un espace dédié à ce type de communication, et de s'assurer des capacités de toutes les parties. À cet

87 Renaissance Numérique (2020), « Réguler les plateformes numériques : Pourquoi ? Comment ? ». Disponible en ligne : https://www.renaissancenumerique.org/publications/reguler-les-plateformes-numeriques-pourquoi-comment

88 Tant par le contenu qu'ils partagent intentionnellement que par les données qu'ils partagent au cours du processus.

89 Voir ses commentaires dans "How Facebook's oversight board could rewrite the rules of the entire internet", *Protocol*, 6 mai 2020 : https://www.protocol.com/facebook-oversight-board-rules-of-the-internet?utm_campaign=The%20Interface&utm_medium=email&utm_source=Revue%20newsletter

égard, les autorités publiques doivent renforcer les capacités des parties prenantes afin de rendre possible la mise en place de processus fonctionnels, collaboratifs et discursifs. Il incombe aux autorités publiques de faciliter la collaboration intra et intersectorielle et le partage de connaissances, de travailler avec la société civile, les chercheurs et les experts techniques pour identifier des méthodes efficaces, et de partager ces méthodes avec tous les acteurs. L'Observatoire européen des médias numériques (*European Digital Media Observatory*), qui prend forme actuellement, est bien adapté à ce rôle au niveau européen. En France, l'Observatoire de la haine en ligne suggéré dans la loi Avia[90] pourrait jouer ce rôle au niveau national. Ce dernier est proposé sous l'autorité du Conseil supérieur de l'audiovisuel (CSA), le régulateur chargé de veiller à la mise en œuvre de la loi. Il convient toutefois de veiller à éviter les redondances et les incohérences entre ces organismes. En outre, les pouvoirs publics doivent veiller à ce que tous les opérateurs de plateformes, en particulier ceux ayant moins de ressources à consacrer à ces questions, soient consultés et pris en compte dans la création de régulations. À cette fin, les autorités de régulation pourraient exiger qu'une analyse d'impact accompagne toute régulation. Elles pourraient également renforcer les capacités techniques des plateformes disposant de ressources moindres : une approche consisterait à allouer les revenus issus des amendes liées aux violations des régulations au renforcement des capacités de ces dernières.

La question de l'interopérabilité, par exemple, est particulièrement importante pour les autorités publiques chargées de vérifier les capacités de toute la gamme des plateformes d'hébergement de contenus générés par les utilisateurs. Alors que les conversations politiques et les décisions techniques commencent à définir les contours de la portabilité et de l'interopérabilité des données, il est crucial que toutes les plateformes soient prises en compte afin d'éviter la création de normes et de protocoles qui en excluent certaines ou les empêchent de bénéficier de ces développements. L'interopérabilité et la portabilité des données visent à permettre aux utili-

90 Consulter la version la plus récente du texte (datant du 13 mai 2020) : http://www.assemblee-nationale.fr/dyn/15/textes/l15t0419_texte-adopte-provisoire.pdf

sateurs de se déplacer plus librement entre les plateformes (et ainsi concrétiser pleinement un droit dont ils bénéficient déjà en vertu du RGPD), et à uniformiser les conditions de concurrence entre les plateformes. Cela ne suffit toutefois pas à établir la concurrence, car ce sont les compétences et les moyens déployés par les plateformes numériques qui sont les facteurs ultimes leur donnant un avantage concurrentiel, et non les données en elles-mêmes. En effet, toutes les plateformes ne favoriseront pas l'interopérabilité (par exemple les opérateurs de niche bien établis) de peur que cette ouverture ne renforce de nouveau les plus grands acteurs ayant la plus grande capacité d'innovation en leur donnant accès à leurs données[91]. Les pouvoirs publics doivent veiller à ce que tous les opérateurs de plateformes pertinents soient entendus dans ce débat, afin d'éviter que l'interopérabilité ne devienne un outil permettant de renforcer de manière disproportionnée le pouvoir des quelques acteurs dominants.

ENTRETENIR UNE CULTURE DE MODÉRATION EN LIEN AVEC LES UTILISATEURS

Une approche collaborative impliquant les utilisateurs peut aider les opérateurs de plateformes à modérer les contenus, en particulier eu égard aux aspects contextuels particulièrement sensibles. Dans le cas de Wikipédia et de Framasoft, deux organisations à but non lucratif, cette approche communautaire s'impose comme une nécessité financière tant que ces dernières continuent à fonctionner à grande l'échelle. Mais les mérites de cette approche ne doivent pas être considérés comme purement financiers. En effet, le fait que la modération communautaire ne soit pas rémunérée suscite des inquiétudes. Une approche collaborative nécessite des processus discursifs, et pas seulement l'externalisation de la main-d'œuvre. Des structures de gouvernance sont alors nécessaires pour faciliter cette participation. Des exemples de modération plus fructueuse et plus inclusive suggèrent que les systèmes les plus efficaces sont multidimensionnels, dans le sens où ils comportent de nombreux niveaux de participation organisés autour d'une équipe de modération centrale, et reposent sur une communication claire et forte entre les différentes couches. Christina Dinar décrit Wikipédia et le site allemand Gutefrage.net comme des systèmes « en oignon », avec leurs structures à nombreuses couches organisées autour d'un noyau central[92]. La modération des contenus étant un défi depuis l'émergence du web moderne, il convient également de tirer les leçons des prémices en matière de gouvernance communautaire (par exemple la modération communautaire semi-transparente dans les communautés de blog comme MetaFilter[93] ou Slashdot).

En France, ONG et militants expriment depuis longtemps la nécessité pour les citoyens de se réapproprier leurs espaces en ligne face aux contenus toxiques par le biais de mobilisations et de contre-discours (SOS Homophobie, SOS Racisme, #StopHateMoney, Projet Seriously). De même, beaucoup déplorent l'« effet spectateur », c'est-à-dire le fait que les utilisateurs soient témoins de contenus toxiques mais ne réagissent pas forcément. De fait, l'engagement des utilisateurs est souvent présenté sous trois angles : 1. le fait d'être un « spectateur actif » ; 2. le fait de signaler officiellement du contenu ou ; 3. le fait de réagir à un contenu toxique. Ce type d'engagement de la part des utilisateurs doit s'inscrire dans un changement de comportement sur les plateformes en ligne bien plus large. Les termes « spectateur actif », « citoyenneté numérique », « résilience de l'utilisateur », souvent évoqués, tendent à faire de l'utilisateur final un acteur principal. La contre-parole en ligne est un phénomène important mais qui, en pratique, ne parvient pas

91 Renaissance Numérique (2020), « Réguler les plateformes numériques : Pourquoi ? Comment ? ». Disponible en ligne : https://www.renaissancenumerique.org/publications/reguler-les-plateformes-numeriques-pourquoi-comment

92 Témoignage de Christina Dinar, ancienne Directrice de projet chez Wikimédia Germany. Entretien du 25 février 2020.

93 Voir : https://metatalk.metafilter.com/24732/Taking-Care-of-a-Fruit-Tree-Moderation-on-Metafilter

toujours à s'élever au-dessus de la mêlée formée par les contenus toxiques[94]. Il est donc nécessaire de renforcer et de formaliser les canaux de contribution des utilisateurs et de créer des mécanismes valorisant ce type de participation.[95] Le renforcement des canaux existants en ce qui concerne les plaintes, les recours et les réparations est par exemple un moyen immédiat pour les plateformes de recadrer la participation des utilisateurs. Toutefois, cette responsabilité en matière de renforcement des capacités est partagée avec les régulateurs. Comme l'a proposé récemment Renaissance Numérique dans une note sur la «plateformisation» de la régulation des services numériques[96], les régulateurs pourraient proposer une approche «macro» de la participation des utilisateurs finaux dans l'ensemble des services numériques au moyen d'une plateforme. Inspiré de la logique des plateformes numériques elles-mêmes, une telle initiative pourrait regrouper des informations, notamment sur des cas litigieux, provenant de toute une série d'opérateurs de plateformes. Ce système permettrait de rationaliser et de structurer les contributions de millions d'utilisateurs finaux, et donnerait du poids à ces derniers dans leur dialogue avec ces services, notamment au travers de la construction d'instruments de régulation et de modération (indicateurs, processus, etc.). Bien entendu, s'il est important pour les opérateurs de plateformes de favoriser le dialogue avec leurs utilisateurs, une approche collaborative ne devrait pas être imposée de manière à contraindre davantage les opérateurs qui sont moins en mesure de le faire. Le fait de confier cette macro-responsabilité aux autorités de régulation permettrait de se prémunir contre ce scénario.

Enfin, les plateformes devraient se charger d'éduquer et d'outiller les utilisateurs dans le cadre de la construction d'une culture de la modération.

94 Renaissance Numérique (2017), «Agir face à la haine sur Internet dans une société collaborative». Disponible en ligne: https://www.renaissancenumerique.org/ckeditor_assets/attachments/184/vnum_note_finale_seriously.pdf

95 Les plateformes pourraient considérer le contre-discours comme une forme de modération du contenu, et explorer les décisions de curation et de conception de contenu pour le faciliter et le promouvoir.

96 Renaissance Numérique (2020), «Réguler les plateformes numériques: Pourquoi? Comment?». Disponible en ligne: https://www.renaissancenumerique.org/ckeditor_assets/attachments/498/note_regulation_des_plateformes.pdf

La modération va bien au-delà de la simple suppression de contenu, et, en ce sens, nécessite souvent une pédagogie à l'égard de l'utilisateur. Cette pédagogie pourrait être intégrée dans la conception et les caractéristiques des plateformes, comme cela est déjà le cas pour certaines[97]. La pédagogie constitue également un mécanisme de recours transparent et efficace: l'existence de politiques claires et sans ambiguïté contribue à réduire la répétition des infractions et à accroître la confiance dans la gouvernance des plateformes. À ce titre, les plateformes se doivent de fournir les ressources nécessaires aux utilisateurs afin de leur permettre de comprendre la logique qui sous-tend les décisions de modération.

97 Par exemple, par le biais d'avertissements automatiques (des messages apparaissant à la vue de l'auteur, lui demandant de réfléchir avant de poster un contenu potentiellement dangereux). Dans le contexte de la pandémie de Covid-19 et de l'«infodémie» qui s'ensuivit, de nombreuses plateformes ont mis à dispositions de leurs utilisateurs du matériel pédagogique et des dispositifs de «coup de pouce» pour lutter contre la désinformation (alertes, redirections vers des sources scientifiquement étayées, etc.)
"How Facebook can Flatten the Curve of the Coronavirus Infodemic", Avaaz, 15 avril 2020: https://secure.avaaz.org/campaign/en/facebook_coronavirus_misinformation/?utm_campaign=The%20Interface&utm_medium=email&utm_source=Revue%20newsletter

CONCLUSION
LA MODÉRATION, UN LEVIER DE LA DÉMOCRATIE

En matière de modération des contenus, il est crucial de valoriser les processus de collaboration et de construire des modèles participatifs, tant dans le cadre des politiques publiques, qu'au-delà. À cet égard, il existe certaines pratiques que la régulation peut encourager, et d'autres qui nécessitent soit la participation de la société civile, soit des actions directes de la part des plateformes. En plus de l'« obligation de diligence » *(duty of care)* des opérateurs de plateformes, il est impératif de mettre en place un système de gouvernance basé sur des processus démocratiques et centrés sur l'utilisateur. Dans le même temps, les responsables politiques se doivent de façonner la régulation de manière à ne pas renforcer la prédominance de quelques plateformes au détriment des autres. Comme cette analyse a tenté de le montrer, bien que les outils et les méthodes des opérateurs « industriels » deviennent souvent la norme dans l'écosystème des plateformes, tout opérateur aurait des leçons à tirer des approches plus « communautaires » et « artisanales » de la modération des contenus. Assurément, des études supplémentaires sont nécessaires, en particulier des analyses comparatives portant sur les approches de la modération des différentes plateformes. Il serait également utile d'analyser plus en détail les typologies présentées ici (industrielle, artisanale et communautaire), par exemple en examinant davantage les pratiques de modération sur les plateformes décentralisées et à code source ouvert, ainsi que sur les plateformes fonctionnant grâce à la publicité ou aux abonnements.

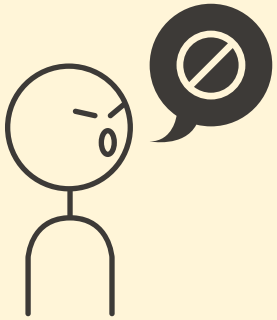
Certains événements récents, de l'affrontement entre le président américain Donald Trump et Twitter, au phénomène d'« infodémie » ayant accompagné la crise du Covid-19, ont rappelé la nécessité de modifier fondamentalement notre façon d'aborder non seulement la modération des contenus, mais également les enjeux démocratiques en question. Ces épisodes soulignent également la nécessité de procéder à ces changements par le biais d'une délibération démocratique et d'une régulation élaborée avec toutes les parties prenantes. Si la régulation est nécessaire pour réduire la toxicité de nos espaces publics en ligne et pour faire respecter les droits fondamentaux et le débat démocratique, elle ne doit pour autant pas devenir un remède pire que la maladie. Une régulation qui encouragerait, directement ou indirectement, les pratiques de modération des contenus

les moins transparentes, les moins responsables et les moins centrées sur l'utilisateur (parmi lesquelles la modération automatisée *ex ante*, le recours à une technologie centralisée et à droit d'accès exclusif, l'utilisation de bases de données non transparentes, etc.) risquerait de nuire non seulement à la qualité de nos espaces en ligne, mais également à leur variété. Ainsi, pour soutenir un large éventail d'expression en ligne, les cadres en matière de régulation doivent veiller à ne pas réduire davantage la diversité des plateformes disponibles. De nouveaux acteurs doivent être en mesure d'entrer sur les marchés, de se développer sur ces derniers et de devenir durables, afin d'offrir aux utilisateurs une diversité d'espaces leur permettant de s'exprimer et de se rassembler, de se sentir en sécurité et d'être entendus. Sans cela, la concentration des plateformes d'hébergement des contenus générés par les utilisateurs, couplée à l'homogénéisation et à l'industrialisation des méthodes de modération des contenus, risque d'accentuer les défis que la modération des contenus cherche à relever en matière d'expression en ligne.

POUR ALLER PLUS LOIN

- Bloch-Wehba, H. "Automation in Moderation", *Cornell International Law Journal*, à paraître (2020)
- « Réguler les plateformes numériques : Pourquoi ? Comment ? », Renaissance Numérique (mai 2020)
- "Nine Principles for Future EU Policy Making on Intermediary Liability", Center for Democracy and Technology (avril 2020)
- « Cyberharcèlement : lecture académique de ce phénomène », Renaissance Numérique (avril 2020)
- "Recommendations on Content Governance", Access Now (mars 2020)
- Douek, E. "The Rise of Content Cartels: Urging transparency and accountability in industry-wide content removal decisions", Knight First Amendment Institute at Columbia University (février 2020)
- "Online Harms White Paper: Initial consultation response", Gov.uk (février 2020)
- Clark, F., Gasser, H., Ross, T. "Content and Conduct: How English Wikipedia Moderates Harmful Speech", The Berkman Klein Center for Internet & Society at Harvard University (décembre 2019)
- Roberts, S. T. "Behind the Screen, Content Moderation in the Shadows of Social Media", Yale University Press (août 2019)
- « Rapport de la mission "Régulation des réseaux sociaux" – Expérimentation Facebook », remis au Secrétaire d'État en charge du numérique (mai 2019)

- “Beyond the ‘Big Three’, Alternative platforms for online hate speech”, The sCAN Project (avril 2019)
- Tworek, H., Leerssen, P. “An Analysis of Germany’s NetzDG Law”, A working paper of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression (avril 2019)
- Newton, C. “The Trauma Floor: The secret lives of Facebook moderators in America”, The Verge (février 2019)
- « Fake News, Faire face aux troubles informationnels à l’ère numérique », Renaissance Numérique (mars 2018)
- Echikson, W., Knodt, O. “Germany’s NetzDG, A key test for combating online hate”, Center for European Policy Studies, CEPS Policy Insight (décembre 2018)
- Caplan, R. “Content or Context Moderation ?: Artisanal, Community-Reliant, and Industrial Approaches”, Data & Society (novembre 2018)
- « Agir face à la haine sur Internet dans une société collaborative », Renaissance Numérique (juillet 2017)
- Grimmelmann, J. “The Virtues of Moderation”, Yale Journal of Law and Technology (2015)



REMERCIEMENTS

Renaissance Numérique a organisé un séminaire sur les enjeux de la modération des contenus toxiques, afin de mieux comprendre les problèmes rencontrés par les « autres » plateformes, c’est-à-dire celles qui dépassent le cadre général des politiques publiques (Twitter, Facebook, Youtube) et qui favorisent des modes de modération plus «artisanaux» ou «communautaires», et de réfléchir ensemble aux pistes d’action possibles. Intitulé *« Comment intégrer l’ensemble des opérateurs de plateformes dans le débat sur la modération ? »*, ce séminaire a eu lieu le 14 février 2020 et a réuni des représentants des plateformes, des membres de la société civile, des chercheurs et des représentants des institutions publiques françaises. Nous tenons à remercier tous les participants à cette matinée d’échanges, ainsi que les personnes interviewées dans le cadre de cette publication.

Robyn Caplan, chercheuse auprès de l’institut de recherche Data & Society et doctorante à l’Université Rutgers, auteure de *Content or Context Moderation?* (2018), a fourni des informations précieuses pour ce projet, du cadrage du séminaire jusqu’à la publication finale. Nous tenons également à remercier particulièrement Christina Dinar, directrice adjointe du Centre for Internet and Human Rights, en Allemagne, pour ses points de vue partagés lors de multiples entretiens et pour sa relecture de la publication finale.

Vous trouverez ci-dessous la liste complète des participants au séminaire du 14 février 2020

- Sarah Durieux, Directrice, Change.org France
- Baltis Mejanes, Cheffe de cabinet et Conseillère parlementaire d’Adrien Taquet
- Charlotte Collonge, Chargée de communication et de contre-discours, Comité interministériel de prévention de la délinquance et de la radicalisation (CIPDR)
- Elise Fajgeles, Policy Officer, Délégation Interministérielle à la Lutte Contre le Racisme, l’Antisémitisme et la Haine anti-LGBT (DILCRAH)
- Lucile Petit, Cheffe du département ‘Services de médias audiovisuels à la demande, Distribution, Nouveaux services’, Conseil supérieur de l’audiovisuel (CSA)
- Salwa Toko, Présidente du Conseil National du Numérique (CNNum)
- Stéphane Koch, Consultant en stratégie digitale
- Clément Reix, Affaires publiques et réglementaires, Dailymotion
- Justine Atlan, Directrice, Association e-Enfance
- Enguerrand Leger, Co-fondateur et Community Manager, GensdeConfiance.fr

- Léo Laugier, Doctorant, Institut Polytechnique de Paris
- Lucien Castex, Secrétaire Général, Internet Society France
- Camille l'Hopitault, Policy Officer, Ligue internationale contre le racisme et l'anti-sémitisme (LICRA)
- Hector de Rivoire, Manager affaires publiques, Microsoft France
- William Schun, Coordonnateur affaires publiques, Microsoft France
- Laure Durand-Viel, Cheffe de projet 'Régulation des plateformes numérique', Ministère de la Culture
- Betty Jeulin, Stagiaire, Pinsent Masons
- Louise Florand, Avocate, Point de Contact
- Iris de Villars, Responsable Desk Technologie, Reporters Sans Frontières
- Benoît Loutrel, Responsable de la mission 'Régulation des réseaux sociaux', Secrétariat d'État au Numérique
- Jean Gonié, Directeur affaires publiques Europe, Snap
- Pauline Birolini, Head of the Legal Department, SOS Racisme
- Valentin Stel, Chargé de projet, SOS Racisme
- Andrea Cairola, Programme Specialist, Division de la liberté d'expression, de la démocratie et de la paix, UNESCO
- Juliette Sénéchal, Maître de conférence, Faculté des sciences juridiques, politiques et sociales, Université de Lille
- Isabelle Jaquemet, Directrice des opérations, Webedia
- Julien Lopez, Advocate General 'Gaming & E-sports', Webedia
- Pierre-Yves Beaudouin, Président, Wikimédia France
- Willie Robert, Vice-Président, Wikimédia France
- Lucien Grandval, Public Policy and Communication Lead, Yubo

Un entretien a été réalisé à la suite du séminaire avec Maïtané Maiwann, Modératrice à l'instance Mastodon de Framasoft, « Framapiaf ».

DIRECTRICE DE LA PUBLICATION

Jennyfer Chrétien, Déléguée générale, Renaissance Numérique

RÉDACTION

Claire Pershan, Chargée de mission, Renaissance Numérique

RENAISSANCE
numerique

À PROPOS
DE RENAISSANCE NUMÉRIQUE

Renaissance Numérique est le principal think tank français indépendant dédié aux enjeux de transformation numérique de la société. Réunissant des universitaires, des associations, des grandes entreprises, des start-ups et des écoles, il vise à élaborer des propositions opérationnelles pour accompagner les acteurs publics, les citoyens et les acteurs économiques dans la construction d'une société numérique inclusive.

Renaissance Numérique
22 bis rue des Taillandiers - 75011 Paris
www.renaissancenumerique.org

Juillet 2020