Publictionnaire
Dictionnaire encyclopédique et critique des Publics

Modération

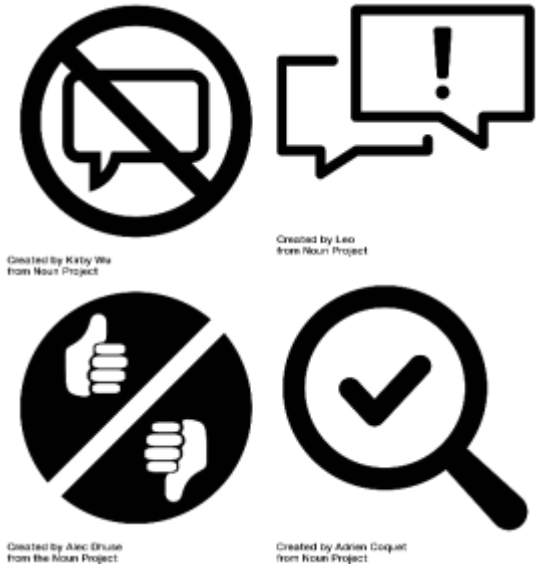
Romain Badouard

Référence électronique
Romain Badouard, Modération. *Publictionnaire. Dictionnaire encyclopédique et critique des publics.* Mis en ligne le 13 octobre 2021. Accès : http://publictionnaire.huma-num.fr/notice/moderation/

Modération

La modération sur le web et les réseaux sociaux est l'activité qui consiste à animer, encadrer, examiner et filtrer les contenus produits par les internautes au sein d'un espace d'échange numérique (discussions, commentaires, publications, etc.). Cette gestion s'appuie généralement sur une charte ou des règles de publication qui fixent les contours de ce qui peut se dire ou non sur le site en question. L'activité des modérateurs et modératrices consiste alors à faire respecter ces règles, *via* différentes méthodes : retrait des contenus prohibés, suppression des comptes qui les publient, mise en quarantaine des auteurs et autrices, invisibilisation des publications, contextualisation des propos tenus, etc.

Si le verbe « modérer », en français, désigne l'action de limiter ou tempérer la virulence d'un propos ou d'un sentiment, son application aux pratiques d'échange revêt également une dimension constructive : le modérateur ou la modératrice est aussi celle ou celui qui organise et valorise les prises de parole du public qui participe à un débat. La modération des contenus, principalement humaine dans les premiers temps du web, tend à s'automatiser par le recours à des outils de détection automatique. Par ailleurs, elle est sujette à controverses, notamment parce qu'elle peut représenter un certain nombre de dangers pour la liberté d'expression des internautes.



Exemples d'icônes correspondant à la modération. Sources : *The Noun Project* (CC BY 3.0). Crédits : *no talking* by Kirby Wu ; *moderation* by Leo ; *Content Moderation* by Alec Dhuse ; *complete* by Adrien Coquet.

L'âge de la modération de masse

À l'époque d'un internet pré-web, puis au sein des communautés en ligne du début des années 1990, la tendance est à l'auto-organisation. Au sein des forums et des listes de

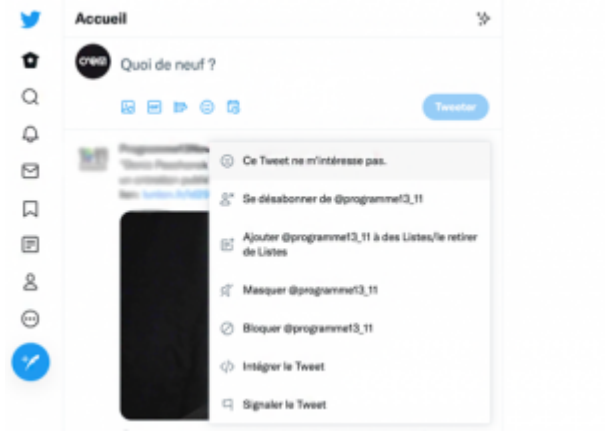
discussion, des utilisateurs volontaires assument le rôle de modérateurs ou de modératrices, animant les débats et rappelant à l'ordre les membres agressifs ou irrespectueux. Informelle, la modération est aussi publique et transparente : les décisions des modérateurs et modératrices se prennent et s'appliquent aux yeux de tous ; elles peuvent être discutées et contestées. En 1995, l'Internet Engineering Task Force, un des organismes en charge du développement des standards techniques de l'internet, publie une charte baptisée « Netiquette » (Hambridge, 1995), qui liste un ensemble de principes à respecter pour garantir des discussions respectueuses dans les espaces d'échange numériques. Cette charte non contraignante constitue dans les années 1990 et 2000 un texte de référence pour les créateurs de forums, qui y puisent un ensemble de principes pour définir les règles de fonctionnement de leurs propres services.

À la fin des années 2000 et au début de années 2010, s'initie un phénomène de « recentralisation » du web autour de quelques services dominants (Smyrnaios, 2017). Les grandes plateformes de réseaux sociaux, comme Facebook et YouTube, passent la barre du milliard d'utilisateurs au milieu des années 2010, devenant de fait les espaces privilégiés du débat public en ligne. Cette centralisation des échanges offre l'opportunité aux pouvoirs publics de faire pression sur ces firmes afin de reprendre la main sur le dossier déjà ancien de la régulation des contenus en ligne (Schafer, 2018) : depuis les années 1990, les tentatives de réglementation de l'internet par les pouvoirs publics en Europe se sont en effet heurtées à un ensemble d'obstacles à la fois techniques, juridiques et culturels. L'échec de ces tentatives a contribué à construire, dans l'opinion, l'image d'un internet impossible à réguler et hermétique aux réglementations nationales en termes de pratiques culturelles comme d'expression publique (Badouard, 2020). De leur côté, les plateformes se sont montrées hostiles à ces tentatives de régulation, se cachant derrière leur posture de « plombiers », gérant des « tuyaux », mais affichant une neutralité radicale par rapport aux contenus qui y circulent.

À la fin des années 2010, cette pression renouvelée des pouvoirs publics pour accentuer les efforts de modération trouve un écho favorable du côté des plateformes, et cela pour différentes raisons. D'abord, les grandes firmes du web vont y trouver un intérêt économique. Tarleton Gillespie (2018) a bien décrit, dans son ouvrage de référence *Custodians of the Internet*, l'évolution du service fourni par les plateformes à cette époque, de la mise à disposition d'outils d'expression pour les internautes, vers la configuration d'espaces de consommation de contenus culturels. Cette évolution, qui impose aux plateformes de garantir à leurs usagers des espaces de consommation standardisés et aseptisés, se double d'une pression des annonceurs eux-mêmes, qui souhaitent que les publicités qu'ils diffusent soient consultées dans des environnements consensuels, où la violence expressive des uns ne vienne pas parasiter l'expérience de consommation des autres. À ces facteurs économiques s'ajoutent des facteurs socio-politiques : à la fin des années 2010, les controverses autour de la propagande terroriste, de la désinformation, des manipulations électorales, du cyberharcèlement et plus généralement de la brutalisation du débat en ligne (Badouard, 2017), ont entaché l'image des plateformes auprès de l'opinion publique. Face à la triple pression des pouvoirs publics, du marché et des usagers, les grandes plateformes vont opérer une réforme en profondeur de leurs politiques de publication et de leurs dispositifs de modération.

La modération en pratique

Les grandes plateformes de réseaux sociaux utilisent trois techniques principales de modération des contenus qu'elles hébergent. La première, et la plus ancienne, est celle du signalement. Elle consiste à mobiliser les internautes eux-mêmes dans la mise en ordre des plateformes en leur permettant de notifier à des évaluateurs professionnels des contenus illégaux ou qui contreviennent aux standards de publication des plateformes. Concrètement, lorsqu'un contenu publié apparaît sur le fil d'actualité d'un internaute, un bouton spécifique lui permet de signaler la publication après l'avoir qualifiée. Le contenu signalé est ensuite transféré à des équipes de modérateurs et modératrices professionnels qui disposent d'un temps généralement réduit pour estimer si la publication contrevient effectivement aux règles de la plateforme. Celle-ci peut alors être supprimée, bloquée, ou voir sa visibilité réduite.



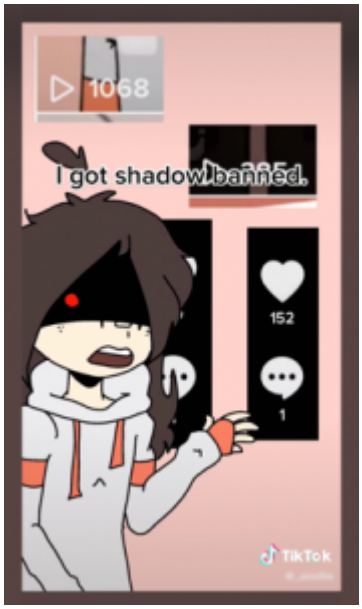
Capture d'écran des différentes propositions pour l'auto-modération sur Twitter. Source : Twitter du Crem.

La place prise par les réseaux sociaux dans le débat public en ligne a produit une véritable industrialisation et professionnalisation de la modération. Le secteur emploierait aujourd'hui plus de 100 000 personnes dans le monde, et serait décliné en différents modèles économiques (Roberts, 2019). Certaines entreprises du web ont des services en interne dédiés à la modération. Il existe également des prestataires externes spécialisés qu'une plateforme peut solliciter pour effectuer des tâches de modération ponctuelles ou permanentes. À ces deux formes d'organisation s'ajoutent les plateformes de micro-travail, qui visent à payer à la tâche des internautes volontaires pour effectuer à distance l'évaluation de contenus. Si les grandes entreprises du web embauchent directement des modérateurs et modératrices, elles font surtout appel à des prestataires externes. On dénombre ainsi 15 000 professionnels travaillant pour Facebook dans le monde, et 10 000 pour YouTube. Cependant, la plupart des plateformes communiquent peu sur leurs dispositifs de modération, à l'instar de Twitter, dont on ne connaît ni le nombre de modérateurs et modératrices ni les modalités d'application précises de ses règles de publication. Cette opacité produit un paradoxe bien décrit par Sarah Roberts (2019) ou T. Gillespie (2018) dans leurs travaux respectifs : plus la modération est nécessaire au service fourni par les plateformes, et moins celle-ci est rendue visible.

À cette première méthode de modération qui consiste à déléguer à des professionnels le travail de régulation des contenus, une seconde méthode employée par les plateformes est

celle de la détection automatique. Face au volume de contenus publiés via leurs services, les plateformes ont opté pour des formes automatisées d'identification des contenus problématiques, en ayant notamment recours à l'intelligence artificielle. Sur Facebook, Instagram, Twitter ou TikTok par exemple, des algorithmes ont pour fonction de scanner les contenus publiés afin de détecter automatiquement les images ou textes qui pourraient contrevenir aux standards de publication. Ces algorithmes sont « entraînés », selon les termes en vigueur dans le domaine de l'intelligence artificielle, sur de grandes bases de données, c'est-à-dire qu'ils vont « se nourrir » des anciens contenus identifiés comme problématiques pour apprendre à les reconnaître. Sur Facebook, la grande majorité de la modération se fait de manière automatique. Dans le domaine de la détection de la nudité ou de contenus violents par exemple, plus de 98 % des contenus contrevenant aux standards de la communauté sont identifiés et retirés de la plateforme directement par des algorithmes, avant même qu'ils ne soient signalés par des internautes.

La dernière méthode revient à jouer sur le degré de visibilité des contenus afin de limiter leur viralité. Plutôt que de supprimer ou bloquer un contenu, l'enjeu est ici d'en paramétrer le public, en limitant le nombre d'internautes qui y sont exposés. Concrètement, il s'agit d'afficher ces contenus plus bas dans les fils d'actualité des internautes, ou de les exclure des recommandations des algorithmes. Moins vus, ces contenus sont moins partagés, et se diffusent moins rapidement. Ces techniques d'invisibilisation, surnommées *shadowban* (« mise au ban ») sur Instagram et TikTok, font preuve d'une certaine efficacité. D'après Google, cette technique permettrait de réduire de 80 % le visionnage de vidéos problématiques. Pour autant, elles sont aussi contestées pour leur opacité, les internautes ne disposant que de très peu d'informations sur les conditions de leur invisibilisation, les critères à partir desquels les décisions sont prises et les voies de recours pour les contester (Badouard, 2020).

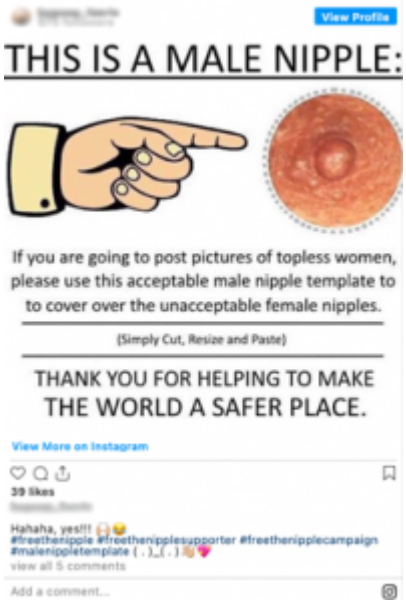


Captures d'écran Tiktok « J'ai été banni ». Source : Tiktok.

Controverses autour de la régulation des contenus

Les pratiques de modération des plateformes sont l'objet de nombreuses controverses. Un premier ensemble de polémiques a trait aux conditions de travail des modérateurs et modératrices. Celles-ci demeurent en effet opaques, les employés s'engageant généralement à la plus stricte confidentialité concernant leurs activités professionnelles. Cependant ces dernières années, plusieurs enquêtes journalistiques ont réussi à percer certains secrets des plateformes concernant leurs pratiques de modération, révélant des cadences de travail infernales, où les modérateurs et modératrices ne bénéficient que de quelques secondes pour statuer sur le caractère délictueux d'une publication. Par ailleurs, certaines plateformes ont recours à la délocalisation de leurs forces de travail, mobilisant dans les pays du sud une main d'œuvre sous-payée pour effectuer la modération dans les pays du nord. À ces questions éthiques s'ajoutent des problématiques d'ordre psychologique : exposés à longueur de journée à des contenus particulièrement éprouvants (viols, meurtres, scènes de torture sur des êtres humains ou des animaux), les anciens employés de ces centres sont nombreux à confier aux journalistes leurs séquelles, certains souffrant même de syndromes de stress post-traumatique.

Un deuxième ensemble de controverses a trait à la manière dont les dispositifs de modération incarnent des formes de domination culturelle, en imposant à l'ensemble des pays dans lesquels sont déployés les services des plateformes une certaine vision de la liberté d'expression et des relations sociales, le plus souvent ancrée dans une culture nord-américaine. La représentation du corps des femmes par exemple, a fait l'objet d'enquêtes approfondies (Myers West, 2017 ; Gillespie, 2018) montrant que les politiques des plateformes en la matière ne sied pas à tous les contextes culturels, et qu'elles font par ailleurs l'objet de contestations de la part des usagers eux-mêmes. De la même façon, la réglementation des discours de haine sur les réseaux sociaux révèlent des différences d'approche fondamentales sur les tensions entre liberté d'expression et lutte contre les discriminations, entre les États-Unis et l'Europe notamment (Girard, 2015 ; Badouard, 2020).



Captures d'écran Instagram

concernant la différence de traitement de modération concernant le corps des femmes et des hommes. Source : Instagram.

Enfin, un troisième ensemble de controverses, peut-être les plus vives de toutes, se concentrent sur la thématique de la privatisation de la censure. À travers la pression effectuée sur les plateformes, les états délégueraient à des firmes privées des fonctions qui étaient auparavant celles des juges, à savoir la capacité à trancher les litiges liés à l'exercice de la liberté d'expression. La suppression des comptes de Donald Trump des principaux réseaux sociaux en janvier 2021, à la suite de l'invasion du Capitole par ses sympathisants, a ainsi fait couler beaucoup d'encre sur le pouvoir dont disposent ces entreprises face à un président élu démocratiquement. Au-delà du cas D. Trump, les formes de censure abusive que produisent au quotidien l'automatisation de la modération et les conditions de travail des modérateurs et modératrices humains, en supprimant des réseaux sociaux des publications légitimes, interrogent les manières de faire respecter les droits fondamentaux des internautes en matière de liberté d'expression en ligne. Dans un certain nombre de pays, comme la France, les pouvoirs publics ont opté pour une régulation par la transparence, qui consiste à superviser le travail de modération des plateformes en exigeant que celles-ci communiquent un certain nombre de données quant à leurs pratiques. Cette régulation par la transparence présente pour autant un certain nombre d'écueils (Badouard, 2021), notamment en ce qu'elle permet aux plateformes de bénéficier d'une « opacité stratégique » (Ananny, Crawford, 2016), en communiquant les données qu'elles souhaitent communiquer, et en gardant secrètes les données sensibles. La possibilité pour des agences indépendantes d'auditer les serveurs des plateformes, afin de certifier les informations transmises par les plateformes aux autorités publiques, constitue aujourd'hui un enjeu essentiel de la régulation des plateformes, notamment dans le domaine de la supervision de leurs pratiques de modération.

Modérer les espaces d'échange en ligne et réguler les prises de parole des publics qui s'y expriment tend à devenir une activité proprement politique, qui implique un certain nombre de responsabilités de la part de celles et ceux qui l'exercent. Le spectre d'une « censure privée », quand la modération est effectuée de manière opaque de la part de plateformes, invite à une forme de contrôle public de ces activités, qui reste encore aujourd'hui à inventer, tant la tâche semble titanesque au vu des volumes de contenus publiés au quotidien.

Bibliographie

Ananny M., Crawford K., 2016, « Seeing without knowing : Limitations of the transparency ideal and its application to algorithmic accountability », *New Media & Society*, 20 (3), pp. 973-989.

Badouard R., 2017, *Le Désenchantement de l'internet. Désinformation, rumeur, propagande*, Limoges, FYP Éd.

Badouard R., 2020, *Les Nouvelles lois du web. Modération et censure*, Paris, Éd. La république des idées/Éd. Le Seuil.

Badouard R., 2021, « Modérer la parole sur les réseaux sociaux. Politiques des plateformes et

régulation des contenus », *Réseaux. Communication, technologie, société*, 225, pp. 87-120.

Gillespie T., 2018, *Custodians of the Internet. Plateforms, content moderation and the hidden decisions that shape social media*, New Haven, Yale University Press.

Girard C., 2015, « Pourquoi punir les discours de haine ? », *Esprit*, 10, pp. 11-22. Accès : https://esprit.presse.fr/article/girard-charles/pourquoi-punir-les-discours-de-haine-38476.

Hambridge S. 1995, « Netiquette Guidelines », *Request for comments*, 1855. Accès : https://datatracker.ietf.org/doc/html/rfc1855.

Myers West S., 2017, « Raging Against the Machine: Network Gatekeeping and Collective Action on Social Media Plaftorms », *Media and Communication*, 5 (3), pp.28-36. Accès : https://www.cogitatiopress.com/mediaandcommunication/article/view/989.

Roberts S. T., 2019, *Behind the Screen. Content Moderation in the Shadows of Social Media*, New Haven, Yale University Press.

Schafer V., 2018, *En construction. La fabrique française d'internet et du Web dans les années 1990*, Bry-sur-Marne, Institut national de l'audiovisuel.

Smyrnaios N.,2017, *Les GAFAM contre l'internet. Une économie politique du numérique*, Bry-sur-Marne, Institut national de l'audiovisuel.