Réseaux sociaux : les nouveaux chemins de la censure

Romain Badouard

Réseaux sociaux : les nouveaux chemins de la censure

Romain Badouard
IFP/CARISM
Université Paris-Panthéon-Assas

Cet article est la version auteur de l'article paru en 2022 dans la revue *Mouvements* sous le titre « Réseaux sociaux : les nouveaux chemins de la censure », *Mouvements*, n°112, 2022, p.137-146.

En juin 2020, un jeune créateur de contenu sur TikTok, Kam Kurosaki[1], fait une expérience déconcertante. Alors que son compte est devenu très populaire sur la plateforme de réseau social à l'occasion du confinement, et que chacune de ses vidéos engrange des centaines de milliers, voire des millions de vues, le jeune états-unien décide de documenter les manifestations du mouvement Black Lives Matter à Los Angeles, quelques semaines après la mort de George Floyd. Une fois rentré chez lui, Kurosaki s'aperçoit que les vidéos qu'il a publiées pendant les manifestations ont été très peu regardées, passant difficilement la barre du millier de vues, soit cent fois moins que ses publications habituelles. Rapidement, il se rend compte qu'il n'est pas seul dans ce cas. Plusieurs créateurs et créatrices sur TikTok font un constat similaire : dès qu'une vidéo relative au mouvement contre les violences policières visant les afro-américains est publiée, son nombre de vues chute drastiquement en comparaison à d'autres types de contenus. L'une de ces créatrices, Onani Banda, constate dans les statistiques de son application que les rares internautes qui continuent de voir ses vidéos sont celles et ceux qui cliquent sur sa page de profil et font donc la démarche volontaire d'aller voir ses vidéos. En revanche, la part d'internautes provenant de la page « *For You* », qui constitue en quelque sorte la page d'accueil de l'application où sont recommandées des vidéos aux usagèr·es, est très faible. Autrement dit, l'algorithme de TikTok semble refuser de recommander les contenus relatifs au mouvement Black Lives Matter à d'autres usagèr·es, condamnant les créateurs et créatrices concerné·es à s'exprimer dans le vide, coupé·es de leur public habituel. Alors que TikTok est l'application la plus téléchargée depuis 2020, et qu'elle constitue dorénavant une plateforme indispensable aux mouvements sociaux qui cherchent à faire connaître leurs actions, ces créateurs et créatrices doivent se rendre à l'évidence : ils et elles ont été *shadow banned*[2].

La doctrine du *shadowbanning*

Le *shadowban*, ou *shadowbanning*, est une technique d'invisibilisation des publications et des comptes utilisée sur les réseaux sociaux afin de limiter l'accès des internautes à certains types de contenus. Grâce à leur position d'infomédiaires[3], c'est-à-dire d'intermédiaires entre les producteurs·ices de contenus et leurs publics, les plateformes comme Facebook, YouTube, Instagram, Twitter ou TikTok sont en mesure de paramétrer très précisément la taille et la nature du public qui est exposé à une publication spécifique. Concrètement, cela revient par

[1] Les témoignages relatés dans cette introduction sont tirés de l'article de Megan McCluskey « These TikTok Creators Say They're Still Being Suppressed for Posting Black Lives Matter Content », *Time*, 22 juillet 2020.
[2] L'entreprise qui possède TikTok justifiera cette censure par invisibilisation par un bug informatique, mais les témoignages récoltés par *Time* laissent penser que le *shadowbanning* des contenus liés au mouvement Black Lives Matter a été récurrent.
[3] F. REBILLARD, N. SMYRNAIOS, « Les infomédiaires, au cœur de la filière de l'information en ligne », *Réseaux*, n° 160-161, 2010, p. 163-194.

exemple à limiter le nombre d'affichages d'un *post* sur les fils d'actualité de Facebook, Instagram ou Twitter, ou à ne plus faire recommander une vidéo particulière par les algorithmes de YouTube ou de TikTok. Les publications visées par une sanction d'invisibilisation ne sont pas supprimées, elles sont toujours accessibles en ligne, mais leur taux d'affichage sur les applications est considérablement réduit. Moins vues, elles sont moins partagées, et leur diffusion s'en trouve ralentie. Le *shadowban* traduit ainsi une nouvelle rationalité d'exercice de la censure sur internet, qui repose non pas sur l'idée d'empêcher une information d'être rendue publique, mais sur celle de noyer cette information sous un flot de contenus tiers afin de la rendre invisible.

Invisibiliser des prises de parole indésirables relève d'une technique ancienne de modération. Dès les débuts de l'internet grand public dans les années 1990 et 2000, les modérateur·rices de forums y ont recours pour gérer les trolls qui viennent pourrir les conversations, en les laissant s'exprimer sans qu'aucun·e autre participant·e ne voit s'afficher leurs interventions sur les fils de discussion. Lassés de voir leurs publications n'occasionner aucune réaction, les trolls finissent par quitter les forums. La pratique du *shadowbanning* reste alors circonscrite à quelques sites et ne constitue qu'une sanction ponctuelle à des comportements jugés antisociaux par les communautés d'internautes. À partir des années 2010, où s'impose la domination des grandes plateformes états-uniennes sur le débat public en ligne et la recentralisation du web autour des services qu'elles proposent, le *shadowbanning* devient une véritable doctrine de régulation des contenus.

Chez Méta par exemple (qui possède Facebook et Instagram), est instaurée en 2016 la politique du « *remove, reduce, inform* », qui veut que les contenus contrevenant aux standards édictés par l'entreprise soient retirés de la plateforme (*remove*), que les usagèr·es soient informé·es de la fiabilité des contenus qu'ils et elles consultent (*inform*) et que les contenus évalués comme étant de mauvaise qualité, mais qui restent conformes aux standards, voient leur visibilité limitée (*reduce*). Une stratégie similaire est mise en place par Google sur YouTube en 2019. La politique dite des « 4Rs », pour « *Remove, Raise, Reward and Reduce* », a pour principe de faire retirer de la plateforme les contenus qui violent les standards (*remove*), d'offrir des bonus de visibilité aux sources jugées fiables (*raise and reward*), tout en limitant la visibilité des contenus de mauvaise qualité (*reduce*). Twitter, entreprise qui communique moins sur ses dispositifs de modération que ses deux principaux concurrents, assume également de limiter la visibilité de certains messages, en jouant sur l'affichage des tweets dans les fils d'actualité, ou en réduisant leurs options de partage (retweets, réponses, etc.). Il en va de même pour TikTok, dont les algorithmes paramètrent les taux d'affichage sur la page d'accueil de l'application (« *For You* ») suivant un certain nombre de critères, et qui peuvent ainsi permettre à une vidéo d'être vue par des millions d'internautes, ou au contraire d'être reléguée dans les limbes du web.

Le *shadowban* s'avère être une technique efficace pour casser la viralité de certains contenus jugés indésirables par les plateformes, mais qui n'enfreignent pas forcément les règles de publication (fausses informations ou sous-entendus racistes par exemple). Elle comporte cependant un écueil important : son opacité. Les internautes qui en font les frais ne disposent en général que de très peu d'informations concernant les sanctions dont ils et elles sont les cibles, et les conditions de sortie d'une mise au ban sont rarement clairement explicitées. Par ailleurs, les lois votées en Europe ces dernières années concernant la régulation des plateformes, qui visent entre autres à exiger d'elles davantage de transparence quant à leurs pratiques de modération, n'abordent pas les enjeux de l'invisibilisation, permettant ainsi aux

grandes firmes de la Silicon Valley de rester très discrètes sur ce sujet[4]. Sur le plan juridique pourtant, filtrer les contenus publiés par les internautes et organiser leur distribution à un public en fonction de différents niveaux de visibilité revient à assumer une fonction proprement éditoriale pour les plateformes. Cette posture est d'autant plus surprenante que dans le long débat concernant leur responsabilité quant aux contenus qu'elles hébergent, les plateformes ont toujours cherché à incarner un rôle d'intermédiaire technique neutre vis-à-vis des publications des internautes. Alors que les journalistes *gatekeepers* (gardien·nes) décidaient, à l'ère de la domination des médias de masse, quelles prises de parole étaient légitimes pour être rendues publiques, à l'époque des plateformes de réseaux sociaux, les grandes firmes du numérique décident de ce qui mérite d'être vu et mis en débat sur le web.

Aux États-Unis, la pratique du *shadowbanning* est éminemment controversée. Si un article de *The Atlantic* en a récemment fait « le problème numéro un des entreprises de la tech[5] », les attaques à son égard viennent principalement des milieux conservateurs, qui se plaignent d'être invisibilisés par des entreprises acquises aux idéaux démocrates. En 2018 par exemple, des élu·es républicain·es avaient accusé Twitter de les invisibiliser en masquant leur compte depuis la barre de recherche du site. Mais c'est surtout la suppression des comptes de Donald Trump par Facebook, YouTube et Twitter lors de l'invasion du Capitole en janvier 2021 qui a suscité l'indignation. Si les incitations à la commission d'actes violents telles que celles proférées par Trump sont effectivement interdites par ces plateformes, la censure par des entreprises privées d'un président élu démocratiquement a suscité des inquiétudes au-delà de ses soutiens habituels. La controverse a rebondi quelques semaines plus tard, lorsqu'un élu républicain de Floride a déposé une proposition de loi visant à interdire, entre autres, les techniques d'invisibilisation. Le *Stop Social Media Censorship Act*, finalement retoqué par la justice[6], prévoyait notamment de rendre illégal pour les plateformes le fait de réduire la visibilité d'un contenu publié par ou sur un·e candidat·e à une élection locale ou à l'échelle de l'État, sans en informer les usagèr·es concerné·es.

L'automatisation de la modération

La généralisation du *shadowbanning* accompagne une autre tendance lourde de la modération sur les réseaux sociaux ces dernières années : son automatisation. Face à la quantité astronomique de contenus publiés tous les jours par les internautes sur ces plateformes, les grandes entreprises de la Silicon Valley ont parié sur l'intelligence artificielle afin de modérer de façon automatique les publications des internautes. Concrètement, des algorithmes sont « entrainés » sur des bases de données contenant des *posts* retirés des plateformes par des modérateur·rices humain·es afin d'apprendre à les reconnaître. Ces mêmes algorithmes vont ensuite passer en revue les nouveaux contenus publiés afin d'y détecter des mots, des tournures de phrases ou des images interdites.

La part des contenus modérés automatiquement ne cesse de progresser par rapport à celle des contenus évalués par des modérateur·rices humain·es. Par exemple, sur Facebook, la proportion de contenus modérés car jugés haineux est passée de 23,6 % au dernier trimestre

[4] R. BADOUARD, « Shadowban. L'invisibilisation des contenus en ligne », *Esprit*, n° 11, 2021, p. 75-83.
[5] G. NICHOLAS, « Shadowbanning Is Big Tech's Big Problem", *The Atlantic*, 28 avril 2022.
[6] De manière qui peut sembler paradoxale, les tribunaux aux Etats-Unis ont tendance à considérer la suppression et le filtrage de publications comme relevant de la liberté d'expression des plateformes. Sur ce sujet, voir P. Auriel et M. Unger, « Les règles de la modération. Débat public, pouvoir privé et censure sur les réseaux sociaux », *Esprit*, n°11, 2021.

2017 à 95,1% au dernier trimestre 2021[7]. Cette proportion a suivi la même courbe sur Instagram et semble concerner tous les types de contenus interdits. Chez Google, YouTube ne propose pas de chiffres aussi détaillés dans ses rapports de transparence mais la PDG de la plateforme, Susan Wojcicki annonçait en décembre 2017 que 98% des vidéos détectées pour « extrémisme violent » au dernier trimestre 2017 l'avaient été par des outils automatisés[8]. Cette proportion dépassait 99% en ce qui concernait les commentaires.

Si la détection automatique s'avère efficace pour modérer certains types de contenus, comme les images de nudité par exemple, elle pose deux problèmes majeurs aux plateformes lorsqu'il s'agit d'évaluer la nocivité de prises de parole publiques : d'une part, il est souvent facile de contourner ces algorithmes, par exemple en attribuant par convention une valeur raciste à un mot anodin (de nombreuses communautés d'extrême-droite utilisent par exemple le mot « suédois » pour désigner ironiquement des personnes originaires de pays arabes ou africains) ; d'autre part, ces algorithmes font des erreurs et produisent de la censure abusive en détectant et retirant des plateformes des contenus légitimes (par exemple quand un internaute dénonce un message raciste en le reproduisant). Pour limiter ces risques de censure abusive, la plupart des grandes plateformes ont mis en place des procédures d'appel qui permettent aux internautes d'exiger une seconde évaluation de leur publication lorsqu'ils estiment son retrait illégitime et, le cas échéant, d'obtenir sa republication.

D'après les chiffres publiés par Facebook[9], au second trimestre 2021, 31,5 millions de contenus ont été supprimés de la plateforme sous prétexte qu'ils constituaient des discours de haine (*hate speech*). Sur l'ensemble de ces suppressions, 1,4 million ont fait l'objet d'une procédure d'appel, procédures qui ont donné lieu à 411 000 republications. Pour le dire autrement, près d'un tiers (29% exactement) des procédures d'appel donnent lieu à une restauration des contenus supprimés. Chez YouTube, les proportions sont similaires[10] : au dernier trimestre 2021, 5,6% des vidéos supprimées ont fait l'objet d'une procédure d'appel, qui ont conduit à des republications dans 20,3% des cas. Suivant les trimestres, le taux de restauration oscille entre 20 et 30%.

La crise sanitaire de 2020 n'a fait qu'accentuer ce processus d'automatisation de la modération. À partir du 16 mars 2020, les modérateurs et modératrices de YouTube ou Facebook ont été confiné·es. Les employé·es étant contraint·es de ne pas communiquer sur leur travail par des accords de confidentialité stricts, et étant donné la nature extrêmement violente de certains contenus, les deux firmes ont décidé que le travail de modération ne pouvait pas être exercé dans un cadre personnel et familial. Elles ont donc confié les clés de leurs dispositifs de modération aux outils de détection automatique. Sur YouTube, pendant cette période, le nombre de suppressions de vidéo a été multiplié par deux, et le taux de restauration après appel est passé à 49,3% au second trimestre 2020.

Ces chiffres démontrent deux choses : d'une part, ils confirment que l'automatisation de la modération s'accompagne de formes de censure abusive, puisque plus la part de contenus modérés automatiquement progresse, plus la part de restauration après appel est importante ;

---

[7] Chiffres tirés des rapports de transparence de Facebook disponibles à l'adresse : https://transparency.fb.com/data/

[8] S. WOJCICKI, « Expanding our work against abuse of our platform », *YouTube Official Blog*, 5 décembre 2017.

[9] Chiffres tirés des rapports de transparence de Facebook disponibles à l'adresse : https://transparency.fb.com/data/

[10] Chiffres tirés des rapports de transparence de YouTube disponibles à l'adresse : https://transparencyreport.google.com/youtube-policy/removals?hl=fr

d'autre part, ils soulignent la nécessité de garantir des procédures d'appel aux internautes afin de limiter les effets liberticides de la modération automatisée. Etant donné les quantités de contenus que doivent traiter les plateformes, il paraît illusoire de prétendre que la modération humaine, mieux à même de prendre en considération le contexte des échanges et donc de limiter les cas de censure abusive, suffira à répondre aux défis posés par la modération de masse. Aujourd'hui, toutes les plateformes ne garantissent pas un droit d'appel, d'autres en revanche vont plus loin encore. Facebook, par exemple, a fait couler beaucoup d'encre en créant en 2020 un Conseil de surveillance (*Oversight Board*), surnommé par la presse Cour Suprême des contenus, qui doit trancher les litiges liés à la modération sur la plateforme, à la manière d'une juridiction de troisième degré[11]. Composé de personnalités de la société civile indépendantes de la firme, le Conseil prend des décisions contraignantes pour l'entreprise, qui s'engage à les mettre en application. Reste que le Conseil en question est financé par Facebook. Si de telles institutions paraissent aujourd'hui nécessaires, leur création pourrait également être prise en charge par les pouvoirs publics ou par des organismes internationaux indépendants. En Europe, de nouvelles législations, notamment en Allemagne, en France et au Royaume-Uni, cherchent à encadrer les pratiques de modération des plateformes, en exigeant de leur part la publication de rapports de transparence. A l'échelle européenne, le Digital Services Act, présenté par la Commission à la fin de l'année 2020, va plus loin en préconisant des audits indépendants des serveurs des grandes plateformes, afin de certifier l'authenticité des données transmises aux régulateurs. Aujourd'hui en effet, rien ne permet de garantir que les informations communiquées par les plateformes sont fiables. Des agences ou institutions publiques ou judiciaires, qui prendraient en charge la gestion des litiges qui persistent entre plateformes et usager.ères au-delà de l'appel, restent encore à inventer.

La modération sur les réseaux sociaux, enjeu de luttes sociales

Les règles de publication des plateformes ne sont pas des lois immuables. Elles évoluent au gré du temps, en fonction des sensibilités de l'époque mais aussi des contestations dont elles sont la cible. Les « standards » ou « règlements de la communauté » sont ainsi l'objet de mobilisations visant à les faire évoluer. Un exemple souvent rappelé dans la littérature académique à ce sujet a trait au mouvement de mères de famille qui se sont mobilisées entre 2007 et 2012 contre la politique de Facebook visant à retirer de sa plateforme les photos d'allaitement[12]. À cette époque, la plateforme supprime en effet toutes les photos de mères donnant le sein sous prétexte de leur « obscénité », motif contesté par les mères en question qui organisent des manifestations et des mobilisations en ligne pour faire entendre leur voix. La compagnie, pourtant, n'en démord pas : toute photo qui laisse entrevoir un téton est supprimée en vertu des politiques de la plateforme concernant la nudité.

C'est dans ce contexte que débute en 2012 le mouvement #FreeTheNipple (« libérez le téton »). Le mouvement vise à susciter une prise de conscience de l'absurdité de certaines règles de modération en maniant habilement l'absurde : des artistes lancent par exemple un concours de retouches de photographies de femmes à la poitrine dénudée en remplaçant leurs tétons par des tétons masculins, qui eux sont autorisés. De nombreuses personnalités du monde du spectacle soutiennent publiquement le mouvement, ce qui lui permet de bénéficier d'une large couverture médiatique et de gagner en popularité. Finalement, en juin 2014, Facebook décide de modifier ses conditions de publication pour autoriser certaines

[11] R. BADOUARD, « Qui contrôle Facebook ? », *AOC*, 8 juin 2021.
[12] Voir par exemple T. GILLESPIE, *Custodians of the Internet: platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press.

photographies montrant des poitrines dénudées, notamment les contenus à visée médicale et celles de femmes allaitant leur enfant.

Les mobilisations des internautes visant à faire évoluer les règles de publication posent la question de la représentation des publics dans la gouvernance des plateformes. La modération sur les réseaux est en effet le théâtre d'un paradoxe évident : alors que les plateformes incitent à l'expression des internautes et qu'elles construisent des modèles économiques extrêmement profitables à partir de la rentabilisation indirecte de ces expressions, les internautes ne disposent d'aucun moyen de faire entendre leur voix quant à la définition des règles de publication. À l'inverse, le web regorge d'espaces collaboratifs qui construisent leur dispositif de modération sur un principe d'auto-organisation. Si Wikipédia en est l'un des meilleurs exemples[13], des forums de discussion populaires comme ceux de Reddit, des réseaux sociaux comme Mastodon ou des plateformes de partage comme Twitch délèguent de diverses manières une partie de la charge de modération aux internautes. Sur Twitch par exemple, s'il existe des règles communes à l'ensemble de la plateforme, les responsables des chaînes décident de règles spécifiques et désignent des modérateurs et modératrices chargé·es d'animer les échanges dans les fils de commentaires. Des outils de modération automatisée sont également mis à disposition des internautes afin de filtrer l'usage de certains mots ou expressions que des communautés souhaitent prohiber, par exemple dans le but de limiter la violence expressive.

Lorsque les internautes ne disposent que de peu de moyens d'action pour influencer la modération à l'intérieur des plateformes, ils et elles tentent de contraindre ces dispositifs en ayant recours à des pressions externes. Au printemps 2021 par exemple, différents mouvements et personnalités féministes avaient porté plainte contre Instagram pour censure. Elles reprochaient notamment au réseau social d'avoir supprimé des *posts* qui comprenaient la phrase « comment faire pour que les hommes arrêtent de violer ? ». Elles demandaient notamment à la justice de contraindre la firme Méta, qui possède Instagram, à rendre publiques un certain nombre d'informations concernant leurs outils de modération sur la plateforme.

Cette revendication d'une transparence accrue est au cœur de nombreuses actions de la société civile. Dans le cas des mouvements LGBTQ+, le chercheur Thibault Grison a par exemple documenté les différentes manières dont des individus et des associations tentaient de faire la lumière sur les formes de censure algorithmique dont ils et elles étaient la cible[14] : rétro-ingénierie pour comprendre le fonctionnement des algorithmes de détection, actions de groupe pour faire appel des décisions des plateformes, mobilisation des médias, appels aux boycotts, etc. Ces actions nous rappellent que ce que nous connaissons aujourd'hui des pratiques de modération des plateformes, nous l'avons appris de travaux de recherche, d'enquêtes journalistiques ou de fuites provenant de lanceurs d'alerte : la transparence n'est jamais venue spontanément des plateformes. Dans un contexte d'opacité généralisée, la production de données alternatives, dans une logique s'apparentant à celle du « statactivisme[15] », permet de saisir les déterminants des dispositifs de modération et de développer des stratégies pour en contourner les abus.

---

[13] D. CARDON, J. LEVREL, « La vigilance participative. Une interprétation de la gouvernance de Wikipédia », *Réseaux*, n° 154, 2009, p. 51-89.

[14] T. GRISON, « The fight against abusive content moderation as a model for new content regulation methods », ICA Pre-conference on *Alternative Content Regulation on Social Media*, 25 mai 2021.

[15] I. BRUNO, E. DIDIER, J. PREVIEUX, *Statactivisme. Comment lutter avec des nombres*, Paris, La Découverte, 2014.

De leur côté, des organismes de recherche indépendants se mobilisent pour produire de nouveaux indicateurs afin d'évaluer les pratiques des plateformes et leur évolution dans le temps. L'organisation états-unienne Ranking Digital Rights s'est par exemple spécialisée dans la production de rapports d'analyse des politiques des géants du numérique en termes de respect des droits de leurs usagèr·es, à travers la publication annuelle du Corporate Accountability Index. La Global Content Governance Survey mesure quant à elle le niveau de confiance des usagèr·es dans les politiques de modération de Facebook et d'Instagram dans 22 pays à partir de sondages réalisés auprès de 6 600 personnes. L'enjeu est ici de déterminer la légitimité des plateformes à décider seules des règles de modération, à les faire appliquer et à organiser des procédures d'appel, afin de nourrir une réflexion plus globale sur des formes de régulation transnationale et démocratique des contenus sur les réseaux sociaux.

Parmi ces initiatives de la société civile, les Santa Clara Principles est peut-être celle qui a eu la portée la plus significative. En mai 2018, une coalition d'universitaires, de militant·es et d'associatifs s'est réunie dans le but d'établir une charte des standards minimum que devraient respecter les plateformes de réseaux sociaux dans leurs pratiques de modération. La coalition a identifié trois principes fondamentaux : *numbers*, c'est-à-dire la publication de chiffres et statistiques relatifs aux *posts* supprimés et comptes suspendus ; *notice*, c'est-à-dire la nécessité de fournir des informations aux internautes dont les *posts* sont supprimés, ainsi qu'à celles et ceux qui les ont signalés, afin de les tenir au courant des procédures d'évaluation dont iels sont l'objet et de leurs résultats ; *appeal*, c'est-à-dire la garantie de procédures d'appel afin que les internautes puissent bénéficier d'un réexamen de leurs publications lorsqu'ils ou elles estiment avoir été injustement censuré·es. Bien que les Santa Clara Principles ne présentent aucun caractère juridiquement contraignant, la charte a été adoptée par un certain nombre d'acteurs de l'économie numérique qui se sont engagés à les respecter, parmi lesquels on retrouve Apple, Facebook, Google, LinkedIn ou Twitter.

Conclusion

Face à la quantité toujours plus importante de contenus publiés tous les jours via leurs services, les grandes plateformes de réseaux sociaux se sont engagées dans deux directions principales pour réformer et adapter leurs dispositifs de modération. D'une part, au-delà de la simple suppression de comptes et de *posts*, elles mettent en œuvre des pratiques de filtrage qui visent à diminuer la visibilité de certains contenus afin de les couper de leur public habituel. D'autre part, elles automatisent la détection de contenus indésirables en pariant sur l'intelligence artificielle. Ces deux pratiques ont pour effet d'accroitre les pratiques de censure abusive sur les plateformes, où des publications légitimes (c'est-à-dire qui n'enfreignent pas les règles de publication) se voient invisibilisées ou supprimées.
Si les plateformes commencent à communiquer sur leurs pratiques de modération, nous ne connaissons rien des profils, notamment politiques, des internautes censurés. En France comme aux Etats-Unis, les milieux conservateurs et l'extrême-droite ont fait de la censure sur les réseaux sociaux un cheval de bataille, accusant des firmes dont les directeur.rices assument publiquement leurs préférences démocrates et progressistes de chercher à les museler. A l'autre bout du spectre politique, des mouvements issus de la gauche radicale portent des accusations similaires, mais en l'absence de contrôle public ou judiciaire des plateformes, nous manquons aujourd'hui de données indépendantes qui permettraient de documenter ces cas de censure politique.
Dans un contexte d'opacité généralisée, la société civile se mobilise en contestant les choix des plateformes et en exigeant davantage de transparence de la part des entreprises qui les

possèdent. Des associations vont plus loin en produisant leurs propres données sur le fonctionnement des algorithmes de détection automatique et sur le fonctionnement du *shadowbanning*, voire en proposant de nouveaux indicateurs visant à évaluer les politiques des plateformes. Cette production militante de données vise notamment à alerter les médias et les pouvoirs publics dans le cadre de nouveaux plans de régulation des plateformes qui se mettent en place aujourd'hui en Europe et dans le monde[16].

En janvier 2018, la chercheuse turco-états-unienne Zeynep Tufekci publiait dans la revue *Wired* un article devenue célèbre sur les nouvelles formes de censure apparaissant sur internet. Dans cet article, elle y interroge un paradoxe apparent : nous vivons aujourd'hui un âge d'or de la liberté d'expression, dans la mesure où il n'a jamais été aussi simple dans l'histoire de prendre la parole et de diffuser ses idées à un public ; et pourtant, jamais dans l'histoire le pouvoir de quelques acteurs privés sur la circulation des idées n'a été aussi important. « Nous n'avons pas à nous résigner », nous dit-elle en conclusion de cet article, « Facebook a 13 ans, Twitter 11 et Google 19. À ce moment de l'histoire de l'industrie automobile, il n'existait ni ceinture de sécurité ni airbags ». De nouvelles règles du jeu peuvent être imposées aux géants du web.

[16] R. BADOUARD, *Les nouvelles lois du web. Modération et censure*, Paris, Le Seuil, 2020.