Évolution de la gouvernance de la liberté d'expression en ligne : Vers une régulation actuarielle de la parole ?

Valentine Crosset, Chaire Good in Tech

Il est aujourd'hui devenu banal de considérer qu'internet a mis à disposition un espace dans lequel peuvent régner des valeurs sociétales, telles que la liberté d'expression. Pendant longtemps, les valeurs du Premier amendement ont largement dominé ces espaces communicationnels. Pourtant, on peut aujourd'hui constater une période de turbulente transformation de la gouvernance de l'expression en ligne. L'approche de laissez-faire et de l'autorégulation ont été vivement critiquées ces dernières années. Cette note propose de faire le point sur cette évolution et de montrer en quoi le contexte de la gouvernance en ligne qui prévaut aujourd'hui est sensiblement différent des années 1990-2000. Elle s'inscrit dans le cadre d'une recherche post-doctorale sur la modération des contenus menée dans le cadre de la Chaire Good in Tech au médialab de Sciences Po.

Les représentations du web sont passées d'un accent mis sur les effets positifs et les promesses d'internet à la mise en évidence des inconvénients et des menaces[1]. Il est vrai qu'au début d'internet, l'utopisme technologique a rapidement célébré sans discernement le rôle d'internet dans les processus démocratiques et de transformations de l'espace public[2]. Or, les environnements numériques peuvent être rudes et tumultueux. De plus en plus, chercheurs et activistes ont montré que certains de ces espaces communicationnels sont devenus la vitrine de discours haineux et extrémistes, du harcèlement, du *doxing*, du *revenge porn* ou encore de la désinformation[3].

Ce passage d'une approche positive à une approche négative de l'espace numérique n'est pas sans conséquence sur les valeurs et logiques qui sous-tendent la liberté d'expression. Il ne s'agit plus de valoriser un discours axé sur les droits individuels (en particulier ceux des utilisateurs), mais de recourir à un discours sur les risques. Ce nouveau type de discours oblige à peser les avantages et les inconvénients systémiques de certains discours et les interventions qui pourraient limiter leurs effets négatifs. Cette évolution montre que si la gouvernance de la liberté d'expression était autrefois dominée par l'approche individualiste

[1] Zittrain, J. L. (2019). Three Eras of Digital Governance. Available at SSN : https://papers.ssrn.com/sol3/papers.cfm?abstract_id=345843.

[2] Voir Benkler, Y. (2009). *La richesse des réseaux : marchés et libertés à l'heure du partage social*). Lyon : Presses Universitaires de Lyon ; Castells, M. (1998). *La société en réseaux*. Paris : Fayard ; Jenkins, H. (2006). *Convegence Culture : Where Old and New Media Collide*. New York, NY : University Press ; Papacharissi, Z. (2002). The virtual sphere: The internet as a public sphere. *New media & society*, *4*(1), 9-27.

[3] Voir Citron, D. K. (2014). *Hate crimes in cyberspace*. Cambridge : Harvard University Press; Citron, D. K., & Franks, M. A. (2014). Criminalizing revenge porn. *Wake Forest L. Rev.*, *49*, 345; Franks, M. A. (2011). Sexual Harassment 2.0. *Md. L. Rev.*, *71*, 655; Marsden, C., Meyer, T., & Brown, I. (2020). Platform values and democratic elections: How can the law regulate digital disinformation?. *Computer Law & Security Review*, *36*, 105-373.

du Premier amendement[4], elle semble s'être aujourd'hui essoufflée. À cet égard, beaucoup de travaux considèrent qu'il est nécessaire de changer de paradigme pour réguler les contenus numériques[5]. En raison du volume des informations rendues publiques par les internautes, les plateformes numériques sont de plus en plus enclines à "fonctionner de manière actuarielle" et à utiliser les probabilités pour réguler leurs flux informationnels[6].

Chacune des conceptions de l'espace numérique rend compte de valeurs importantes qui "peuvent à la fois renforcer et limiter la liberté d'action individuelle, y compris celle de s'engager dans des comportements nuisibles"[7]. L'objectif aujourd'hui est de retracer les différents cadres de valeurs concurrents en ce qui concerne la liberté d'expression en ligne. Nous verrons ainsi que la gouvernance de la parole en ligne est exploration, jeu infini d'essais et d'erreurs rectifiées.

L'ère du libéralisme

La gouvernance de l'expression en ligne a dans un premier temps promu un laissez-faire avoué à l'égard des contenus publiés par les utilisateurs. Ce mode de gouvernance reflétait, pour partie, les préceptes révolutionnaires cyberlibertariens d'un internet "ouvert" et "libre" en matière d'expression. Il s'agissait alors de transposer dans le domaine de l'information les principes d'une « éthique libertaire classique »[8]. En 1996, John Perry Barlow, parolier des *Grateful Dead* et membre fondateur de *l'Electronic Frontier Fondation*, présentait à Davos sa « déclaration d'indépendance du cyberespace », dans laquelle il déclarait que « nous créons un monde où chacun, où il se trouve, peut exprimer ses idées, aussi singulières qu'elles puissent être, sans craindre d'être réduit au silence ou à une norme »[9]. De tels récits libertaires sont par ailleurs emblématiques des principes libéraux de la jurisprudence constitutionnelle américaine[10]. C'est pourquoi cette forme de libertarisme doit être entendue dans le contexte de la culture politique américaine, qui accorde notamment à la liberté d'expression une valeur cardinale.

---

[4] Ammori, M. (2013). The new New York Times: Free speech lawyering in the age of Google and Twitter. *Harv. L. Rev.*, *127*, 2259; Douek, E. (2021). Governing online speech. *Columbia Law Review*, *121*(3), 759-834; Balkin, J. M. (2018). Free speech is a triangle. *Colum. L. Rev.*, *118*, 2011.

[5] Citron, D. K., & Franks, M. A. (2020). The Internet as a Speech Machine and Other Myths Confounding Section 230 Reform. *U. Chi. Legal F.*, 45 ; Citron, D. K., & Wittes, B. (2017). The internet will not break: Denying bad samaritans sec. 230 immunity. *Fordham L. Rev.*, *86*, 401 ; Franks, M. A. (2019). Fearless Speech. First Amendment Law Review, vol. 17 (294). Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3363020; Wu, T. (2017). *Is the First Amendment Obsolete*. New York: Knight First Amendment Institute, Columbia University. Available at: https://knightcolumbia.org/content/tim-wu-first-amendment-obsolete.

[6] Ananny, M. (2019). Probably speech, maybe free: Toward a probabilistic understanding of online expression and platform governance. Knight First Amendment Institute. Available at: https://knightcolumbia.org/content/tim-wu-first-amendment-obsolete

[7] Zittrain, J. L., *op. cit.*

[8] *Ibid*.

[9] Barlow, J. P. (2000). Déclaration d'indépendance du cyberespace. Dans O. Blondeau (ed.), Libres enfants du savoir numérique (pp. 47-54). Paris : Éditions de l'Éclat.

[10] Loveluck, B. (2015). Internet, une société contre l'État?. *Réseaux,* (4), 235-270.

Pour les plus enthousiastes, internet allait donc inaugurer une nouvelle ère de démocratie culturelle dans laquelle les utilisateurs seraient d'actifs participants à l'espace public[11]. Ces prédictions reposaient sur l'hypothèse qu'internet pourrait perturber le pouvoir des *gatekeepers* traditionnels – notamment les géants des médias et toutes sortes de pouvoirs centralisés. Cette pensée a eu pour conséquence d'affirmer un droit individuel fort : celui de s'exprimer sans subir d'interférences arbitraires. Internet est ainsi rapidement devenu l'emblème d'un nouveau marché des idées, dans lequel l'information circulerait librement sans avoir à prendre en considération d'autres intérêts, notamment ceux des personnes ou groupes qui peuvent être visés par les propos exprimés. C'est dans ce contexte qu'on peut aisément associer cette période à celle d'une ère des « droits individuels » de la gouvernance d'internet[12].

Cet idéal d'un internet "ouvert" et "libre" a été renforcé par un ensemble de lois visant à créer des immunités et des protections pour les hébergeurs de contenus, de façon à faciliter l'innovation et la libre circulation de l'information[13]. C'est notamment le cas de la section 230 du Communications Decency Act de 1996[14], qui a grandement favorisé cette évolution. Cette loi accorde aux propriétaires des plateformes une large immunité en matière de responsabilité sur ce que leurs utilisateurs publient. Sa particularité est d'agir à un double niveau : si la loi n'oblige pas les plateformes à surveiller ce que leurs utilisateurs publient, elle les autorise en même temps à définir, de façon privée, leurs propres règles relatives au contenu publié par les utilisateurs[15].

Il importe toutefois de préciser que l'approche n'a jamais été celle de l'absolutisme de la liberté d'expression[16]. En réalité, les plateformes numériques ont rapidement maintenu une ambivalence entre la crainte d'intervenir et celle de ne pas intervenir dans la définition des règles privées qu'elles imposent à leurs utilisateurs sous le nom de "Charte de la communauté". Même si la liberté d'expression était exaltée, au fur et à mesure de l'émergence de services d'hébergement, plusieurs d'entre elles ont mis en place des « directives communautaires » ou des conditions d'utilisations qui interdisaient certains types de contenus, généralement la pornographie, les discours haineux, l'obscénité et les activités illégales[17]. Il n'était donc pas question pour ces dernières d'ignorer les dangers que peuvent induire certaines expressions. Mais, les valeurs de la liberté d'expression dans le droit constitutionnel américain se sont reflétées sans conteste dans les premières règles privées des plateformes[18]. La modération répondait à un principe d'intervention minimale, quant aux

[11] Voir Balkin, J. M. (2013). Old-school/new-school speech regulation. *Harv. L. Rev.*, *127*, 2296. ; Balkin, J. M. (2017). Digital speech and democratic culture: A theory of freedom of expression for the information society. In *Popular Culture and Law* (pp. 437-494). Routledge ; Benkler, Y., *op. cit.*

[12] Zittrain, J. L., *op. cit.*

[13] Cohen, J. E. (2019). *Between truth and power: The legal constructions of informational capitalism*. Oxford : Oxford University Press.

[14] Communications Decency Act of 1996, 47 U.S.C. § 230. Available at: https://www.law.cornell.edu/uscode/text/47/230

[15] Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. New Haven, CT : Yale University Press.

[16] Ammori, M. *op. cit*. ; Douek, E. *op. cit.*

[17] Gillespie, *op. cit.*

[18] Voir Van Zuylen-Wood, S. (2019, 26 février). "Men are scum": Inside Facebook's war on hate speech, https://www.vanityfair.com/news/2019/02/men-are-scum-inside-facebook-war-on-hate-speech

préoccupations relatives au harcèlement et aux discours haineux en ligne, celles-ci n'atteignaient que rarement le niveau de débats publics ou politiques[19].

L'ère du risque

Durant les vingt premières années de l'internet populaire, l'idéal d'une parole "ouverte" et "libre" est resté la norme dominante[20]. Les choses ont néanmoins profondément changé au cours des dix dernières années. Il ne s'agit plus seulement de valoriser le droit individuel des utilisateurs, comme dans la première période de la gouvernance d'internet, mais de favoriser un environnement d'expression en ligne "sain" et "sûr"[21]. Qu'est-ce qui explique ce changement de paradigme dans l'appréhension des droits et libertés sur internet ? Avant tout, rappelons que depuis quelques années maintenant, internet est plus que jamais devenu un agglomérat d'entreprises privées, si bien qu'elles dominent actuellement le contrôle des flux informationnels[22]. Amazon, Microsoft, Apple et Facebook figurent parmi les principales entreprises, aux côtés de Netflix et des géants chinois Alibaba et Tencent. Les milliards d'utilisateurs qui utilisent internet sont par conséquent capturés par un petit nombre d'entreprises. Par exemple, Facebook contrôle 80% du marché des services sociaux, avec plus de deux milliards d'utilisateurs mensuels dans le monde[23]. La publicité en ligne est quant à elle monopolisée à plus de 60% par Facebook et Google.

Dès lors, cet oligopole, ainsi que d'autres hébergeurs de contenus, comme Twitter et Reddit, sont devenus de fait les arbitres du contenu publié sur le web[24]. C'est en partie en raison de la concurrence et de la pression du marché que les plateformes ont de plus en plus cherché à créer des environnements sains et sûrs. Les plateformes ont effectivement tout intérêt à présenter leur meilleur visage aux nouveaux utilisateurs, aux annonceurs et à leurs partenaires, ainsi qu'au grand public[25]. De plus, elles doivent éviter de perdre des utilisateurs qui seraient victimes de harcèlement, tout comme de faire l'objet de mesures légales lorsqu'elles n'ont pas respecté les lois de certains pays en matière de contenus haineux ou autre[26].

Dans ce contexte, l'autorégulation et la libre circulation de l'information promue par les pionniers prend aujourd'hui une autre tournure. Bien que les plateformes restent attachées au principe de la liberté d'expression, elles sont également devenues plus attentives aux contenus. Selon Tim Wu, trois raisons expliquent ce changement[27]. Pour commencer, il y a eu une réévaluation culturelle plus large concernant la responsabilité des plateformes et la liberté d'expression : les activistes de la dernière décennie ont notamment exposé les effets nuisibles

[19] Gillespie, *op. cit.*
[20] Wu, T. (2019). Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Ordering Systems. *Columbia Law Review*, *119*(7), 2001-2028.
[21] *Ibid.*
[22] Van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford : Oxford University Press.
[23] *Ibid.*
[24] Sylvain, O. (2019). Recovering Tech's Humanity. *Columbia Law Review*, *119*(7), 252-282.
[25] Gillespie, *op. cit.*
[26] *Ibid.*
[27] Wu, T. *op. cit.*

que peuvent comporter certains actes expressifs sur des groupes historiquement marginalisés. Ce danger et ces risques ont d'autant plus été mis en évidence ces dernières années, face à la circulation en ligne d'une grande quantité de discours toxiques. Dans ce contexte, les appels se sont multipliés pour que les plateformes limitent la publication de certaines expressions. Il s'est dès lors opéré un changement de paradigme en matière de liberté d'expression en ligne. Une partie de la société milite effectivement à ce que certaines expressions soient restreintes, pour une meilleure reconnaissance d'autres libertés ou intérêts sociaux.

La deuxième raison pour Tim Wu est plus politique. Les plateformes ont maintes fois été accusées d'avoir toléré la diffusion de discours haineux et de propagande terroriste, ainsi qu'une ingérence étrangère dans les élections. Cette préoccupation a essentiellement émergé à la suite des élections américaines de 2016. Par ailleurs, l'auteur rappelle qu'en dehors des USA, Facebook a fait face à de multiples accusations. Il a notamment été soutenu que la plateforme était utilisée pour organiser et promouvoir la violence dans des pays comme le Myanmar, le Sri Lanka et l'Inde. Cette seconde préoccupation ne résulte pas seulement de la crainte de nuire à des groupes marginalisés, mais aussi de l'effet que peuvent avoir certains discours sur l'équilibre démocratique et géopolitique.

Enfin, la dernière raison pour l'auteur est la consolidation d'un nombre limité de plateformes (Twitter, Facebook et Google) qui impose une conception globale de l'ensemble du réseau et de la distribution des flux informationnels. En 2016, les grandes plateformes numériques ont commencé à prendre au sérieux certains abus en ligne (étant des acteurs privés, ces plateformes ne sont pas limitées par les normes constitutionnelles). Pour ces dernières, il n'était plus tenable d'évoquer la rhétorique du premier amendement, à savoir que la meilleure réponse à un discours malveillant est plus de discours. Peu à peu, les principales plateformes ont décidé de traiter les discours haineux comme potentiellement "violent" et "nuisible" et qui peuvent faire l'objet d'une suspension. Ainsi, les puissants intermédiaires décident désormais des nombreuses catégories de discours qui doivent faire l'objet d'une suppression. Tim Wu observe que celles-ci vont des plus faciles à définir (vidéos de tentatives de suicide, pornographie enfantine) aux plus ambiguës (discours de haine, discours déshumanisant, apologie du terrorisme, glorification de la violence).

Ce mélange de prise de conscience nouvelle et d'état de fait a inévitablement conduit les plateformes à ne plus penser la modération des contenus à travers une lentille individualiste de la liberté d'expression, typique du premier amendement[28]. Les plateformes procèdent maintenant de plus en plus à la mise en balance de la liberté d'expression avec d'autres droits et intérêts. Dans ce contexte, les règles sont rédigées de façon à englober de multiples intérêts et non plus seulement des droits individuels[29]. Cet accent mis sur la proportionnalité et la mise en balance implique que la modération des contenus doit maintenant trouver un équilibre entre la liberté d'expression et d'autres libertés ou intérêts sociaux[30].

[28] Douek, E. *op. cit.*
[29] *Ibid*.
[30] *Ibid.*

L'ère de la statistique

Ces nouvelles exigences qui visent à créer un espace sécurisé opèrent une prise immédiate : elles exigent que les plateformes modèrent plus et plus rapidement. D'où un ensemble de réformes adoptées dans l'urgence, au coup par coup, avec le souci constant de tenir juridiquement responsables les plateformes des réseaux sociaux de ce que leurs utilisateurs publient en ligne[31]. Ce faisant, les questions de management efficace, d'optimisation des moyens et de gestion des ressources humaines ont largement été mises de l'avant-plan[32]. Pour autant, les plateformes ne se heurtent pas seulement à des problèmes de valeurs, mais aussi à des difficultés logistiques conséquentes.

De manière générale, le changement d'échelle apporté par les plateformes marque un tournant dans l'appréhension de la gouvernance de la parole. Quotidiennement, c'est une immense quantité de contenus qui doit être examinée. Depuis fin 2020, Facebook rassemble plus de 2.85 milliards d'utilisateurs actifs. Au premier trimestre de 2021, Facebook a supprimé plus de 242 millions de contenus et 905 millions de spams[33]. YouTube, dont pas moins de 500 heures de vidéos sont téléchargées chaque minute sur la plateforme, a suspendu plus de 9 569 641 vidéos et 1 032 365 719 commentaires entre janvier 2021 et mars 2021[34]. Durant le premier trimestre de 2020, Twitter a pour sa part supprimé 1 927 062 contenus et suspendu 9 524 744 comptes[35]. Ces chiffres ne reflètent que les cas pour lesquels des mesures ont été prises. Comme le rappelle l'ancien responsable de la sécurité de Facebook, Alex Stamos, le nombre total des décisions, qui comprend aussi les contenus pour lesquelles aucune sanction n'a été prise, est beaucoup plus élevé[36].

Dans ce contexte, le débat suggère que les plateformes ont créé une échelle de discours qu'elles ne sont pas en mesure de gouverner de manière fiable[37]. La vitesse et l'échelle d'un ensemble d'utilisateurs mondialisés, qui distribuent en temps réel du contenu, rend impossible d'obtenir une modération correcte[38]. L'erreur est inévitable. À cet égard, Monika Bickert de Facebook, reconnaît qu'une « entreprise qui examine cent mille contenus par jour et maintient un taux de précision de 99% peut encore avoir jusqu'à mille erreurs ». Concernant les vidéos

---

[31] Par exemple, on trouve les nouvelles législations qui ont émergé en Allemagne (la loi NetzDG en 2017), en France (la loi contre la manipulations de l'information en 2018), ainsi que le nouveau règlement européen relatif à la lutte contre la diffusion de contenus à caractère terroriste (2018), ou encore les projets législatifs au Royaume Unis qui concernent la lutte contre la violence en ligne.

[32] Gillespie, *op. cit*. ; Roberts, S. T. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, CT : Yale University Press.

[33] *Community Standards Enforcement Report*, Rapport de transparence Facebook, https://transparency.fb.com/data/community-standards-enforcement

[34] *Application du règlement de la communauté YouTube*, Rapport de transparence YouTube, https://transparencyreport.google.com/youtube-policy/removals?hl=fr

[35] *Application des règles*, Rapport de transparence Twitter, https://transparency.twitter.com/fr.html

[36] Alexander Stamos, Prepared Written Testimony and Statement for the Record of Alexander Stamos, before U.S. House of Representatives Committee on Homeland Security Hearing on "Artificial Intelligence and Counterterrorism: Possibilites and Limitations, 25 juin 2019, https://perma.cc/GRW8-VKPJ

[37] Ananny, *op. cit* ; Douek, *op. cit* ; Masnick, M. (2019, 20 novembre). Masnick's Impossibility Theorem: Content Moderation At Scale Is Impossible To Do Well. *Techdirt,* https://www.techdirt.com/articles/20191111/23032743367/masnicks-impossibility-

[38] Monika Bickert, *Defining the Boundaries of Free Speech on Social Media,* Dans Lee C. Bollinger et Geoffrey R. Stone (eds), The Free Speech Century (pp. 254-271). Oxford : Oxford University Press.

de terrorisme, Google a assuré que les signalements des Trusted Flagger sont exacts dans 90% des cas[39]. Pour réduire le taux d'erreur, le programme cherche à inclure toujours plus d'organisations et de signaleurs de confiance.

Compte tenu de l'échelle, le mieux que les plateformes puissent faire est de minimiser le taux d'erreur[40]. Dans un premier temps, les entreprises ont essentiellement embauché plus de personnes dans leur équipe de sécurité. En 2019, Facebook a indiqué que 35 000 personnes travaillent sur la modération des contenus et à améliorer la sécurité du réseau[41]. Concernant Twitter, ce chiffre s'élève à 15 000 personnes[42]. En 2018, Google souhaitait faire monter à 10 000 le nombre d'employés chargés de lutter contre les contenus susceptibles de violer leurs politiques[43]. Néanmoins, la capacité de surveiller et de faire respecter les normes a radicalement changé au cours des dernières années, en raison de l'augmentation de la « modération algorithmique »[44]. C'est parce que le contrôle ne peut se construire uniquement avec des humains et parce que les règles et politiques laissant encore trop de liberté, que les plateformes ont massivement investi dans les algorithmes, et notamment l'intelligence artificielle.

Ce faisant, la plupart des compagnies des réseaux sociaux ont mis en place des pratiques de retrait automatisées et semi-automatisées[45]. Ce système de modération hybride, composé d'algorithmes et d'humains, fait reposer l'adjudication des droits de la parole sur un terrain statistique mouvant. Comme Mike Ananny l'indique, « la modération des contenus est… probabiliste. Il s'agit d'une confluence de probabilités : un filtre algorithmique a-t-il déclenché un seuil de calcul pour bloquer le contenu offensant, un nombre suffisant d'utilisateurs au cours d'une période donnée ont-ils signalé une quantité suffisante de contenu pour entraîner la suspension d'un compte, et des modérateurs de contenus tiers ont-ils appliqué de manière égale les normes du contenu des plateformes ? »[46]. En cela, l'auteur stipule que les plateformes ne font que rendre probable la circulation de la parole. Dit autrement, les interdictions sont probabilistes et jamais binaires. Pour appuyer ce propos, Mike Ananny donne deux exemples évocateurs[47].

Il rappelle d'abord que les interdictions ne s'appliquent que dans les langues et régions surveillées. Alors que Facebook propose son interface en 111 langues, ses algorithmes ne peuvent détecter les discours haineux dans seulement 30 langues et la propagande terroriste

[39] Walker, K. (2017, Juin 18). *Four steps we're taking today to fight terrorism online*. *Google Blog*. https://blog.google/around-the-globe/google-europe/four-steps-were-taking-today-fight-online-terror/
[40] Douek, *op. cit.*
[41] Voir tweet Facebook, https://twitter.com/facebook/status/1186324897321996288?lang=fr
[42]Szadwoski, M. (2021, 4 février). Pourquoi la modération de Twitter fonctionne si mal. *Le Monde*, https://www.lemonde.fr/pixels/article/2021/02/04/pourquoi-la-moderation-de-twitter-fonctionne-si-mal_6068758_4408996.html.
[43] Levin, S. (2017, 5 décembre). Google to hire thousands of moderators after outcry over YouTube abuse videos. *The Guardian*, https://www.theguardian.com/technology/2017/dec/04/google-youtube-hire-moderators-child-abuse-videos
[44]Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 2053951719897945.; Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2), 2053951720943234.
[45] Sylvain, O. *op. cit*.
[46] Ananny, M. *op. cit.*
[47] *Ibid.*

dans 19 langues[48]. De plus, aucune plateforme n'a traduit ses normes dans toutes les langues que propose son interface. En somme, les interdictions ne sont actuellement pas en mesure de couvrir toutes les langues et les pays du monde. Le bannissement d'Alex Jones est pour l'auteur un autre exemple d'un tel management probabiliste. Au moment des discussions internes concernant la suspension sur Instagram d'Alex Jones, figure de proue du complotisme extrémiste de l'alt-right américaine, Facebook a utilisé un raisonnement actuariel pour défendre son inaction initiale[49]. L'équipe de modération a constaté que le compte d'Alex Jones n'atteignait pas le seuil de préjudice établi par la plateforme. Pour qu'un compte soit supprimé, il doit avoir au moins 30% de contenus en violation à un moment donné. Concernant les commentaires sur le compte d'Alex Jones, ceux-ci n'arrivent pas au seuil de suppression : seuls 23 (4%) des 530 commentaires étaient en infraction.

Si une gestion probabiliste de la gouvernance de la parole semble être un compromis imparfait, mais pragmatique[50], les plateformes révèlent des distributions de probabilité très inégales. Comme le rappelle Mike Ananny, bien que les erreurs statistiques soient faibles, elles sont vécues par des populations particulières et des individus singuliers[51]. Qui supporte la haine non sanctionnée, lorsque Facebook déclare que le compte Instagram d'Alex Jones n'est pas assez offensant pour être sanctionné ? Ou, lorsque Monika Bickert affirme que Facebook maintient un taux de précision de 99% (chiffre difficilement vérifiable), qui vit avec le taux d'erreurs restant ?

Remarques conclusives et recommandations

Si dans un premier temps, la gouvernance de la liberté d'expression a mis l'emphase sur les droits individuels, elle adopte aujourd'hui, dans une logique d'équilibrage, un langage actuariel (le risque et la probabilité) appliqué à des populations et des traces numériques. Une telle reconnaissance, qui prône une approche systémique, semble en contradiction avec le cadre de conception individualiste qui découle de la tradition du Premier amendement. Mais, elle ne peut être d'emblée exclue. Les plateformes numériques existent aujourd'hui à une échelle qui rend la supervision humaine pratiquement impossible. L'accent est dès lors mis sur la recherche de techniques algorithmiques pour diminuer les effets nuisibles de certains discours, non pas dans le but de les éradiquer totalement, mais dans celui de les maintenir à un taux acceptable. Mais, ces techniques sont à ce jour peu transparentes. En réalité, nous n'avons que peu de connaissances sur la manière dont ces algorithmes de modération fonctionnent ni sur les données à partir desquelles ils sont entraînés.

Ce nouveau calibrage apparaît comme une des questions politiques majeures de ce moment clé de la gouvernance de l'expression en ligne. Soumettre les droits d'expression en ligne à

[48] Flick, M. & Paresh, D. (2019, 23 avril). Facebook's flood of languages leave it struggling to monitor content. *Reuters*, https://www.reuters.com/article/us-facebook-languages-insight-idUSKCN1RZ0DW
[49] Kanter, J. (2019, 28 mars). Leaked emails reveal Facebook's intense internal discussion over Alex Jones' 'anti-Semitic' post on Instagram. *Insider*, https://www.insider.com/facebook-emails-reveal-discussion-about-alex-jones-instagram-account-2019-3.
[50] Douek, *op. cit.*
[51] Ananny, M. *op. cit.*

une analyse de proportionnalité ou à une optimisation probabiliste revient à se demander à quelles finalités sociales elles concourent. Autant la liberté d'expression pouvait être intelligible dans le cadre d'une approche libérale, autant logique actuarielle, surveillance algorithmique ou neutralisation des contenus, semblent aujourd'hui moins en phase avec le projet initial de la liberté d'expression en ligne. Par conséquent, c'est la question de la légitimité des plateformes, qui se pose, et à sa capacité très relative de répondre à la demande des autorités et du public de garantir un espace à la fois sûr et sain. Dans le contexte d'une gouvernance privatisée de la parole en ligne, sur quelle base les plateformes établissent-elles les avantages et inconvénients de certains discours? Comment le risque est-il distribué ? Qui négocie les seuils de préjudice ? À partir de quel seuil un discours est-il jugé comme tolérable ou non ? Comment les modérateurs sont-ils formés à reconnaître et résoudre les erreurs ?

Face à ce changement de paradigme dans l'appréhension de l'expression en ligne et afin de promouvoir une gouvernance responsable, il importe à ce stade de favoriser plusieurs pistes. Premièrement, mener des recherches plus approfondies sur la régulation probabiliste des contenus en ligne. L'objectif sera de mieux comprendre les limites et les forces de ce type de gouvernance. Deuxièmement, encourager une plus grande collaboration entre le monde de l'industrie, les acteurs judiciaires, les chercheurs en sciences sociales et en *machine learning*, pour étudier, prévenir et atténuer les risques associés à la modération algorithmique et probabiliste. Troisièmement, identifier les meilleures pratiques concernant la modération de contenus de manière à répondre aux préoccupations en matière de régulation des contenus. Cet échange de bonnes pratiques est une façon d'encourager une gouvernance responsable au sein des plateformes existantes. Quatrièmement, assurer une plus grande transparence de la part des plateformes quant à la précision et aux erreurs des algorithmes de modération, aux garanties et aux procédures si le système défaille, à la manière dont le système déjoue les biais culturels et historiques, comme le racisme et le sexisme. Enfin, mettre en place de certifications et d'audits en ce qui concerne ces systèmes algorithmiques. Il s'agirait d'évaluer de manière indépendante les risques et mesurer l'impact social potentiel d'un tel système sur la liberté d'expression.

Pour aller plus loin :

- Un livre important sur le monde de la modération des plateformes des grands réseaux sociaux, ses enjeux et ses limites: *Gillespie, T. (2018). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. New Haven, CT : Yale University Press.*

- Pour une réflexion sur l'application d'un cadre probabiliste aux questions relatives à la gouvernance de l'expression en ligne : *Ananny, M. (2019). Probably speech, maybe free: Toward a probabilistic understanding of online expression and platform governance. Knight First Amendment Institute. Accessible en ligne : https://knightcolumbia.org/content/tim-wu-first-amendment-obsolete*

- Sur l'idée d'une approche de la gouvernance de l'expression en ligne fondée sur les principes du Premier amendement à une approche d'équilibrage systémique entre la liberté d'expression et d'autres libertés : *Douek, E. (2021). Governing online speech. Columbia Law Review, 121(3), 759-834.*

- Sur l’idée la gouvernance privatisée des contenus et des plateformes numériques: *Klonick, K. (2017). The new governors: The people, rules, and processes governing online speech. Harv. L. Rev., 131, 1598*

- Pour un aperçu de l’évolution des normes d’expression en ligne et sur l'hybridation humain-machine dans la modération des contenus : *Wu, T. (2019). Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Ordering Systems. Columbia Law Review, 119(7), 2001-2028.*