SHADOW BAN

L’invisibilisation des contenus en ligne

Romain Badouard

Éditions Esprit | « Esprit »

2021/11 Novembre | pages 75 à 83

ISSN 0014-0759
ISBN 9782372341905
DOI 10.3917/espri.2111.0075

Shadow ban

L'invisibilisation des contenus en ligne

Romain Badouard

La prolifération des fausses informations et des discours de haine sur Internet pose un problème d'un nouveau genre aux grandes plateformes de réseaux sociaux. Si Facebook, Google, Twitter et leurs concurrents ont, depuis leur création, édicté des règles de publication, délimitant ce qui peut se dire ou non au sein des espaces d'échange qu'elles mettent à disposition des usagers, les *fake news* et propos haineux occupent généralement une zone grise particulièrement difficile à modérer. Les publications qui relèvent de ces deux catégories constituent en effet, pour les plateformes, des contenus de mauvaise qualité, dont la présence sur leurs réseaux est jugée indésirable. Lorsqu'ils sont explicites, ils sont facilement identifiables et peuvent faire l'objet de mesures de retrait ou de mise en quarantaine. Mais dans la pratique, les propos haineux s'expriment généralement à couvert : leurs auteurs utilisent un mot à la place d'un autre pour désigner une cible, manient le sous-entendu et l'ironie, ou ont recours à des symboles racistes, antisémites, homophobes ou misogynes. Les fausses informations, quant à elles, relèvent moins souvent d'une manipulation manifeste que de propos biaisés, décontextualisés, voire absurdes, dont la dimension malveillante n'est pas toujours évidente. Dans les deux cas cependant, ces publications placent les plateformes face à un dilemme : si elles les laissent en ligne, elles sont accusées de laxisme ; si elles suppriment des contenus qui ne relèvent pas directement de propos répréhensibles, elles sont accusées de censure abusive.

La solution expérimentée, puis généralisée, par les principales plateformes à partir de la fin des années 2010 consiste à rendre invisibles les publications problématiques, stratégie communément appelée *shadow ban* (« mise au ban fantôme ») au sein des communautés d'usagers. Son

principe : réduire l'affichage des contenus jugés de mauvaise qualité[1], sans les supprimer, afin de limiter leur visibilité par les internautes. Moins vus, les contenus sont moins partagés et leur diffusion s'en trouve considérablement ralentie. Concrètement, une publication identifiée par un modérateur ou un algorithme comme relevant d'une fausse information ou d'un propos nuisible se verra attribuer une mauvaise note, qui aura pour conséquence de l'afficher plus bas dans les fils d'actualité de Facebook, Instagram et Twitter, ou de moins la recommander aux usagers sur YouTube ou TikTok. Le *shadow ban* consiste ainsi à exercer une forme de régulation des contenus, non pas en faisant pression sur leurs producteurs, mais en configurant leur réception par un paramétrage précis du public qui y sera exposé.

Si la technique s'avère efficace pour modérer les contenus problématiques (d'après YouTube, par exemple, elle permet de réduire le visionnage de contenus complotistes de 80 %[2]), elle soulève des inquiétudes légitimes en matière de censure abusive des espaces de débat en ligne. D'une part, les décisions d'invisibilisation sont prises en toute opacité par des acteurs privés, qui exercent ainsi un véritable pouvoir politique. Nombreuses sont les organisations ou personnalités de la société civile à s'être plaintes, depuis la fin des années 2010, de formes d'invisibilisation arbitraires. D'autre part, les internautes qui font les frais d'un *shadow ban* sont rarement avertis de la sanction dont ils sont la cible : ils continuent de s'exprimer « dans le vide », sans même s'en rendre compte, coupés de leur public habituel. Invisible, offrant peu de prise à la contestation, le *shadow ban* constitue une pièce d'un puzzle bien plus large, celui du cloisonnement de l'espace public en ligne, où la configuration des espaces de débat est dictée par les préférences de chacun, mais aussi par les choix des entreprises qui détiennent les infrastructures informationnelles qui leur confèrent le pouvoir de décider ce qui est vu et débattu à l'échelle d'Internet.

1 - Les contenus jugés de mauvaise qualité ne concernent pas uniquement les fausses informations et les discours de haine, mais également toutes les publications qui se placent à la frontière des interdits définis par les normes de publication, sans les enfreindre directement, et peuvent avoir trait à la publicité masquée, la promotion de la pornographie, l'incitation à la violence, etc.
2 - Voir The YouTube Team, "Our ongoing work to tackle hate" [en ligne], *YouTube Official Blog*, 5 juin 2019.

(In)visibilité des paroles en ligne

Le principe même d'invisibiliser des prises de parole indésirables sur les espaces d'échange en ligne n'a rien de nouveau. Dès les années 1980, avant même l'invention du *web*, les modérateurs disposent, avec les *bulletin board systems*[3], de moyens techniques permettant de limiter l'accès de certains utilisateurs aux fils d'échange. Dans les années 1990 et 2000, l'invisibilisation des internautes sur les forums apparaît comme un moyen efficace de gérer les trolls qui viennent pourrir les conversations, en les laissant s'exprimer, mais sans qu'aucun autre participant ne soit exposé à leurs prises de parole. L'expression consacrée est alors de dire que tel utilisateur est « parti au couvent ». La technique s'avère rapidement efficace : lassés de voir leurs invectives et interventions sur les forums demeurer sans suite, les trolls les délaissent pour d'autres cibles.

Avec le phénomène de recentralisation d'Internet au cours des années 2010, qui voit le débat en ligne, auparavant dispersé entre une multitude de sites et de forums, se concentrer autour des principales plateformes de réseau social, le *shadow ban* devient une véritable doctrine de régulation des contenus. Chez Facebook par exemple (qui possède également Instagram), est instaurée en 2016 la politique du *"remove, reduce, inform"*, qui veut que les contenus contrevenant aux normes édictées par l'entreprise soient retirés de la plateforme *(remove)*, que les usagers soient informés de la fiabilité des contenus qu'ils consultent *(inform)* et que les contenus évalués comme étant de mauvaise qualité, mais qui restent conformes aux normes, voient leur visibilité limitée *(reduce)*. Une stratégie similaire est mise en place par Google sur YouTube en 2019. La politique dite des « 4 R », pour *"remove, raise, reward and reduce"*, a pour principe de faire retirer de la plateforme les contenus qui violent les normes *(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.).

3 - Les *bulletin board systems* (BBS) permettaient d'héberger, *via* des serveurs, des conversations ou des échanges de fichiers sur des réseaux de communication (Internet ou autres) dès les années 1970.

La généralisation du *shadow ban* pose la question de la légitimité des plateformes à décider de ce qui mérite d'être vu et discuté. L'invisibilisation participe en effet d'une nouvelle logique de gouvernement des paroles publiques, avec laquelle les grandes entreprises du numérique disposent d'un pouvoir sans précédent sur l'organisation du débat sur Internet. Dès le début des années 2010, le sociologue Dominique Cardon avait identifié comment le principe de visibilité tendait à remplacer celui de publicité dans l'espace public en ligne[4]. À l'ère des médias de masse, dit Cardon, les informations étaient privées par défaut et devenaient publiques par un tri préalable à la publication effectué par les *gatekeepers* que sont les journalistes, les éditeurs, les producteurs et les programmateurs. Sur Internet, à l'inverse, les informations sont publiques par défaut, et le contrôle éditorial consiste à organiser la visibilité des informations après leur publication, entre une petite minorité qui sera portée à la connaissance des internautes et l'écrasante majorité qui sera bannie dans les limbes d'Internet. Ce contrôle des informations n'échoit plus aux *gatekeepers* traditionnels, mais aux grandes compagnies d'Internet qui gèrent les infrastructures informationnelles, comme les moteurs de recherche ou les réseaux sociaux.

La généralisation du *shadow ban* pose la question de la légitimité des plateformes à décider de ce qui mérite d'être vu et discuté.

Les règles du jeu de ce nouvel espace public ont très tôt été intégrées par les producteurs de contenu, qui, *via* les techniques de *search engine optimization* (SEO), ont cherché à obtenir de meilleurs référencements sur les moteurs de recherche, afin de toucher un public plus large, ou par des entreprises cherchant à recruter un maximum de clients potentiels sur les réseaux sociaux *(growth hacking)*. Les mobilisations politiques en ligne reposent également sur ce principe de visibilité : tout l'enjeu pour un « entrepreneur de cause » va être d'occuper l'espace du débat au moyen de différentes techniques et ressources d'optimisation, quitte parfois à faire taire un opposant *via* le cyberharcèlement militant afin de limiter la visibilité d'arguments contradictoires. Les États eux-mêmes se sont saisis de techniques similaires. En Chine, par exemple, les paroles dissidentes ne sont pas supprimées des réseaux, mais noyées sous un flot

4 - Voir Dominique Cardon, *La Démocratie Internet. Promesses et limites*, Paris, Seuil, coll. « La République des idées », 2010.

artificiel de paroles en faveur du régime afin de simuler un mouvement d'opinion qui lui est favorable *(astroturfing)*. Un récent rapport de l'Irsem a mis en lumière la manière dont les techniques de l'État chinois n'étaient plus uniquement dévolues au contrôle de l'espace public intérieur, mais utilisées comme des armes diplomatiques à destination de la diaspora, des médias et internautes étrangers. Le rapport en question indique ainsi que les différentes commissions des affaires cyber et les bureaux de propagande du pays emploieraient deux millions de commentateurs rémunérés à plein temps, dont le travail consiste notamment à alimenter les espaces de débat en ligne afin d'assurer la promotion de la Chine à l'étranger ou d'éteindre les polémiques la concernant. À ces travailleurs du clic s'ajouteraient vingt millions de « trolls à temps partiel », recrutés pour des opérations ponctuelles auprès d'étudiants chinois, mais aussi *via* l'externalisation de ces opérations à des plateformes de microtravail situées en Malaisie. Les auteurs du rapport en concluent que la Chine s'inspire dorénavant des méthodes de la Russie, visant à assurer son influence par des méthodes d'intoxication et d'intimidation délibérées au sein des espaces d'échange en ligne[5].

Le *shadow ban* constitue ainsi un révélateur parmi d'autres des possibilités de manipulation du débat public en ligne, en jouant notamment sur les architectures informationnelles, c'est-à-dire les algorithmes et les applications qui assurent la promotion d'un sujet plutôt qu'un autre, et qui déterminent la visibilité d'une publication ou la portée d'un message.

Invisibilisation et légitimité démocratique

En Occident, les questions soulevées par la mise en (in)visibilité des informations sur Internet ont moins été appréhendées à travers le prisme de la propagande qu'à travers celui de la censure. À l'été 2018, aux États-Unis, des élus républicains s'étaient émus que certaines figures du parti soient subitement introuvables sur Twitter, accusant le réseau social d'exercer un *shadow ban* à l'encontre des conservateurs pour mettre en lumière les publications des Démocrates. Donald Trump s'était lui-même saisi de l'affaire, accusant Twitter de pratiques illégales et discriminatoires. L'été

5 - Voir Paul Charon et Jean-Baptiste Jeangène Vilmer, *Les Opérations d'influence chinoises. Un moment machiavélien* [en ligne], Institut de recherche stratégique de l'École militaire, septembre 2021.

suivant, en France, à l'autre extrémité de l'échiquier politique, plusieurs pages Facebook de mouvements politiques issus de la gauche radicale avaient dénoncé une chute inexpliquée de leur audience, certaines pages voyant les visionnages de leurs *posts* divisés par cent, voire par mille, sans autres formes de justification de la part du réseau social. Leur participation active au mouvement des Gilets jaunes semblait être, selon les intéressés, la cause de cette censure arbitraire. Plus récemment, au printemps 2021, plusieurs associations féministes françaises ont assigné Instagram en justice pour avoir invisibilisé les comptes de militantes qui avaient affiché la phrase « Comment faire pour que les hommes arrêtent de violer ? ». Elles exigeaient du réseau social qu'il s'explique sur ses pratiques de modération asymétriques, nombre de publications misogynes ayant par ailleurs droit de cité sur la plateforme.

Les exemples d'invisibilisation à visée politique ne manquent pas et appellent à une forme de contrôle démocratique du travail de modération des plateformes[6]. Si la question de la régulation des réseaux sociaux suscite l'intérêt de l'opinion et des décideurs ces dernières années, force est de constater que le sujet de l'invisibilisation constitue l'un des angles morts des lois passées ou en préparation en Europe. En France, la loi sur les manipulations de l'information, entrée en vigueur en décembre 2018, attribue au Conseil supérieur de l'audiovisuel (CSA) un pouvoir de contrôle des activités de modération des plateformes. Dans la pratique, ce pouvoir s'exprime par une simple demande d'accès à un certain nombre d'informations, que les plateformes doivent transmettre tous les six mois *via* des rapports d'activité. Les questionnaires publiés par le CSA à destination des plateformes ne comprennent pas directement de questions relatives aux techniques d'invisibilisation. Ils comportent en revanche des éléments relatifs à la transparence des algorithmes de classement des informations et des dispositifs de modération. Dans leurs réponses, qui sont rendues publiques sur le site du CSA, les plateformes manient habilement le flou. Facebook n'aborde même pas le sujet de sa politique d'invisibilisation, et si YouTube le fait, aucun chiffre, aucune donnée précise ne sont fournis à l'appui, qui permettraient de quantifier ou

6 - Voir Romain Badouard, *Les Nouvelles Lois du web. Modération et censure*, Paris, Seuil, coll. « La République des idées », 2020.

d'objectiver l'ampleur du phénomène[7]. Cet exemple illustre les lacunes de la régulation par la transparence[8], qui offre aux grandes entreprises d'Internet une opportunité d'« opacité stratégique », en se montrant transparentes sur des sujets anodins et secrètes sur des points beaucoup plus sensibles.

Cette négligence des autorités publiques est d'autant plus surprenante qu'en opérant un tel tri entre les informations et en attribuant délibérément des scores de visibilité aux publications des internautes, les plateformes sortent de leur apparente neutralité à l'égard des contenus publiés *via* leurs services. Elles qui, historiquement, ont toujours défendu leur statut d'hébergeur, qui les prémunit de toute responsabilité juridique quant aux publications mises en ligne par leurs usagers, se livrent ici à des activités typiquement éditoriales, qui leur imposent une responsabilité par rapport à ces mêmes contenus. Cette évolution majeure de l'activité des plateformes, de la mise à disposition d'outils d'expression vers la configuration d'espaces de consommation d'informations, constitue un levier sur lequel les autorités pourraient faire pression pour exiger davantage des géants du numérique en termes de transparence et de responsabilité.

Plus surprenant, certaines plateformes adoptent cette évolution vers des activités éditoriales en faisant valoir leur liberté d'expression quant au tri qu'elles opèrent dans les publications des internautes. Dans un article paru dans la *Revue des droits et libertés fondamentaux*, Pierre Auriel et Mathilde Unger analysent le cas de l'affaire Zhang *vs* Baidu. Aux États-Unis, des militants chinois en faveur de la démocratie ont assigné en justice le moteur de recherche Baidu pour avoir rendu leurs publications invisibles. La défense de l'entreprise avait alors consisté à faire valoir que le référencement des informations pouvait s'apparenter à un discours politique à part entière et qu'à ce titre, il devait être protégé par le premier amendement de la Constitution américaine. Le tribunal américain avait alors donné raison au moteur de recherche, arguant que *« trier et présenter revient à exprimer son opinion sur ce qui importe »*. Dans ce contexte, la décision de

7 - Voir *Lutte contre la diffusion des fausses informations sur les plateformes en ligne : bilan de l'application et de l'effectivité des mesures mises en œuvre par les opérateurs en 2019* [en ligne], Conseil supérieur de l'audiovisuel, juillet 2020.

8 - Voir R. Badouard, « Modérer la parole sur les réseaux sociaux. Politiques des plateformes et régulation des contenus », *Réseaux*, n° 225, 2021, p. 87-120 et « Ce que peut l'État face aux plateformes », *Pouvoirs*, n° 177, 2021, p. 49-58.

Baidu de censurer les discours en faveur de la démocratie était elle-même protégée par *« l'idéal démocratique de la liberté d'expression*[9] *»*.

Si une telle défense paraît difficilement envisageable en Europe, les grandes plateformes d'Internet ont compris que la légitimité de leur travail d'invisibilisation reposait sur son externalisation. Facebook ou YouTube ont ainsi délégué à des acteurs tiers le travail d'évaluation de la qualité des publications afin de ne pas être accusés de censure arbitraire. Facebook, par exemple, fait appel à des journalistes professionnels pour évaluer la fiabilité des informations *via* son programme *third-party fact-checking* lancé en 2016. Le principe de ce partenariat est d'embaucher au sein de rédactions reconnues des journalistes membres de l'International Fact-Checking Network, afin d'évaluer des publications qui leur sont signalées en leur attribuant une note de fiabilité. La note en question déterminera ensuite la visibilité des publications évaluées sur les fils d'actualité des internautes. Google, de son côté, fait appel à des internautes ordinaires, réunis au sein de panels, pour évaluer leur expérience de navigation et de consommation de contenus lorsqu'ils utilisent les produits de l'entreprise. Ces *quality raters* doivent notamment se prononcer sur la fiabilité des pages qu'ils consultent sur le moteur de recherche ou des vidéos qu'ils visionnent sur YouTube. Cette évaluation collective influencera par la suite le référencement des sites et la recommandation des vidéos.

La généralisation du *shadow ban* préfigure un débat en ligne cloisonné, stratifié, où l'incertitude règne quant aux informations auxquelles sont exposés nos interlocuteurs. Sur Facebook ou YouTube, l'invisibilisation ne concerne pas uniquement le référencement et les recommandations de *posts* et de vidéos, mais également la gestion des commentaires : les propriétaires de pages ou de chaînes peuvent décider de masquer aux autres participants les interventions d'un internaute en particulier. Comment dès lors qualifier un échange où tous les intervenants ne sont exposés ni aux mêmes prises de parole ni aux mêmes informations, et

9 - Pierre Auriel et Mathilde Unger, « La modération par les plateformes porte-t-elle atteinte à la liberté d'expression ? Réflexions à partir des approches états-uniennes (*Zhang v. Baidu.com*, 2014) et italienne (*CasaPound contro Facebook*, 2019) » [en ligne], *Revue des droits et libertés fondamentaux*, n° 80, 2020.

où les discours accessibles sont des discours autorisés ? Le débat est-il encore *public* sur les réseaux sociaux ? Au début des années 2010, puis à l'occasion de la campagne présidentielle américaine de 2016, a resurgi la controverse autour des « bulles cognitives » ou « informationnelles » sur Internet, les moteurs de recherche et les réseaux sociaux étant accusés de distribuer des informations aux internautes en fonction de leurs préférences personnelles, les enfermant ainsi dans des sphères hermétiques et idéologiquement homogènes. Cet équipement technologique des préférences a pour conséquence de refermer le débat public en ligne, quand la principale contribution d'Internet au fonctionnement des démocraties a été d'ouvrir ce débat à une multitude de nouvelles voix. Avec le *shadow ban*, l'enjeu est dorénavant de savoir si nous devons confier les clés de cet espace public à des pouvoirs privés qui échappent encore largement à toute forme de contrôle démocratique.