Shadowban : l'invisibilisation des contenus en ligne

Romain Badouard

Modération : la politique du *shadowban*

Romain Badouard

Cet article est la version auteur de l'article paru en 2021 dans la revue *Esprit* (n°479) sous le titre « Shadowban. L'invisibilisation des contenus en ligne », pp.75-83.

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 proposent à leurs 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. Pour autant, elles n'enfreignent pas forcément leurs règlements. Les discours de haine, par exemple, sont facilement condamnables lorsqu'ils sont explicites, mais dans la pratique, ces derniers 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 qui par convention sont associés à des prises de position racistes, antisémites, homophobes ou misogynes. La modération des fausses informations, quant à elle, impose aux plateformes de s'ériger en arbitre de la vérité, alors même qu'aucun règlement de réseau social n'interdit le mensonge, le canular ou l'affabulation. 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 sur le dossier épineux des « pollutions informationnelles »[1] ; 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 est celle de l'invisibilisation des publications problématiques, appelée communément *shadowban* au sein des communautés d'usagers. Son principe : limiter la visibilité des contenus de mauvaise qualité, sans les supprimer, afin que ceux-ci soient moins vus par les internautes, donc moins partagés, et ainsi réduire leur viralité. 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 *shadowban* consiste ainsi a exercé une forme de régulation des contenus, non pas en menaçant de sanctions leurs producteurs, mais en configurant leur réception par un paramétrage très 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 la visionnage de contenus complotistes de 80%[2]), elle soulève des inquiétudes légitimes en termes 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 via cette pratique un véritable pouvoir politique (nombreuses sont les organisations ou personnalités

[1] Voir C. Wardle, H. Derakshan, 2017, "Information disorder : Toward an interdisciplinary framework for research and policy making", *Rapport au Conseil de l'Europe*, DGI(2017)09.
[2] Google, 2019, « Our ongoing work to tackle hate », *YouTube Official Blog*, 05/06/2019, disponible en ligne.

de la société civile à s'être plaint, depuis la fin des années 2010, de formes d'invisibilisation arbitraires). D'autre part, les internautes qui font les frais d'un *shadowban* 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 *shadowban* constitue une pièce d'un puzzle bien plus large, celui du cloisonnement de l'espace public en ligne, où les préférences de chacun dictent la configuration des espaces de débat, et où les entreprises qui détiennent les infrastructures informationnelles disposent du pouvoir de décider ce qui est vu et débattu à l'échelle du web.

(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 avec les *bulletin board systems*, avant même l'invention du web, les modérateurs disposent 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 demeurées sans suite, les trolls les délaissent pour d'autres cibles.

Avec le phénomène de recentralisation du web au cours des années 2010, qui veut que le débat en ligne, auparavant dispersé entre une multitude de sites et de forums, se concentre dorénavant autour des principales plateformes de réseau social, le *shadowban* 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 au standards soient retirés de la plateforme (*remove*), que les usagers soient informés de la fiabilité des contenus qu'ils consultent (*inform*) et que les contenus 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.).

La consécration des techniques d'invisibilisation comme dispositif d'organisation du débat public participe à une nouvelle logique de gouvernement des prises de parole 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[3]. A l'ère des médias de masse, nous dit Cardon, les informations étaient privées par défaut, et devenaient publiques

[3] Voir D. Cardon, 2010, *La démocratie internet. Promesses et limites*, Le Seuil, coll. La République des Idées.

par un tri pré-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 du web. Ce contrôle des informations n'échoit plus aux *gatekeepers* traditionnels, mais aux grandes compagnies du web qui gèrent les infrastructures informationnelles, comme les moteurs de recherche ou les réseaux sociaux.

Dans ce nouvel espace de discussion connecté, on passe ainsi d'un régime de gouvernement de la parole publique qui reposait sur une distinction entre dicible et indicible (ce que l'on peut dire ou non), dont la responsabilité incombait à l'auteur, reposant sur des éléments de loi définis à priori et dont la gestion était déléguée à la justice ; à un régime basé sur une distinction entre visible et invisible (ce que l'on peut voir ou non), relevant d'un paramétrage des publics (qui est exposé à l'information, et comment), qui repose sur des évaluations contextuelles (au cas par cas), dont la gestion incombe aux opérateurs privés de plateforme et à leurs partenaires de modération.

Ces nouvelles règles du jeu 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 via différentes techniques et ressources d'optimisation, quitte parfois à faire taire un opposant afin de limiter sa propre visibilité via le cyberharcèlement militant. 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 de paroles pro-régimes afin de simuler un mouvement d'opinion qui lui est favorable (*astroturfing*). Le *shadowban* intègre ainsi un répertoire de procédures et de dispositifs beaucoup plus vaste dans cette bataille incessante pour la visibilité.

Invisibilisation et légitimité démocratique

La généralisation de l'invisibilisation comme technique de gestion des espaces de débat en ligne soulève donc un enjeu majeur de légitimité démocratique. Comment s'assurer que les plateformes ayant recours à ces techniques les mettent en œuvre dans une optique d'apaisement du débat, ou d'optimisation de la qualité des informations qu'elles hébergent, et non dans une logique de censure politique ? A l'été 2018, aux Etats-Unis, des élus républicains s'étaient émus que certaines figures du parti soient subitement introuvables sur Twitter, accusant le réseau social d'exercer un *shadowban* à l'encontre des conservateurs pour mettre en lumière les publications des démocrates. Donald Trump lui-même s'était saisi de l'affaire, accusant Twitter de pratiques illégales et discriminatoires. L'été suivant, en France, à l'autre extrémité de l'échiquier politique, plusieurs pages Facebook de mouvements politiques issues de la gauche radicale avaient dénoncé une chute inexpliquée de leur audience, certaines pages voyant les visionnages de leurs *posts* divisés par 100, voire par 1000, 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 posté 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, nombres de publications misogynes ayant par ailleurs droit de citer sur la plateforme.

Les exemples d'invisibilisation à visée politique ne manque pas, et appellent à une forme de contrôle démocratique du travail de modération des plateformes[4]. 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 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 n'est fourni à l'appui, qui permettrait de quantifier ou d'objectiver l'ampleur du phénomène[5]. Cet exemple illustre les lacunes de la régulation par la transparence, que j'ai déjà eu l'occasion d'aborder à d'autres reprises[6], et qui offre aux grandes firmes du web 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 postés 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 embrassent 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*[7], Pierre Auriel et Mathilde Unger relatent le cas de l'affaire Zang VS Baïdu. Aux Etats-Unis, des militants

[4] Voir R. Badouard, 2020, *Les nouvelles lois du web. Modération et censure*, Le Seuil, coll. La République des Idées.
[5] « Lutte contre les infox : le CSA publie son premier bilan », CSA.fr, 30 juillet 2020.
[6] Voir R. Badouard, 2021, « Modérer la parole sur les réseaux sociaux. Politiques des plateformes et régulation des contenus », *Réseaux*, n°225, p.87-120 ; R. Badouard, 2021, « Ce que peut l'état face aux plateformes », *Pouvoirs*, n°177, p.49-58.
[7] P. Auriel et M. Unger, 2020, « La modération par les plateformes porte-t-elle atteinte à la liberté d'expression ? Réflexion à partir des approches états-unienne et italienne », *Revue des droits et libertés fondamentaux*, n°80, disponible en ligne.

chinois pro-démocratie ont assigné en justice le moteur de recherche Baïdu pour avoir invisibilisé leurs publications. 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 rechercher, arguant que « trier et présenter revient à exprimer son opinion sur ce qui importe ». Dans ce contexte, la décision de Baïdu de censurer les discours pro-démocratie était elle-même protégée « par l'idéal démocratique de la liberté d'expression ».

Si une telle défense paraît difficilement envisageable en Europe, les grandes plateformes du web 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 accusé 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 la firme. 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.

Conclusion

Au-delà du contrôle démocratique des techniques d'invisibilisation, la généralisation du *shadowban* 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 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, qui veut que les moteurs de recherche et les réseaux sociaux distribuent des informations aux internautes en fonction de leurs préférences personnelles, les enfermant ainsi dans des sphères hermétiques et idéologiquement homogènes. Avec le *shadowban*, cette logique poursuit son extension pour s'insinuer dans les conversations du quotidien.