Modérer la parole sur les réseaux sociaux. Politiques des plateformes et régulation des contenus

Romain Badouard

Modérer la parole sur les réseaux sociaux : politiques des plateformes et régulation des contenus

Cet article est la version auteur de l'article suivant paru dans la revue *Réseaux* en 2021 : « Modérer la parole sur les réseaux sociaux : politiques des plateformes et régulation des contenus », *Réseaux*, n°225, 202, p. 87-120.

Romain Badouard
IFP / CARISM
Université Paris 2 Panthéon-Assas

Depuis l'élection présidentielle américaine de 2016, un tournant semble s'être opéré dans les relations entre états et plateformes du web. A la suite des auditions par le Congrès des responsables des principaux réseaux sociaux dans le cadre de l' « affaire russe » et de l'affaire « Cambridge Analytica », ces derniers ont entrepris une réforme profonde de leurs politiques de modération et de régulation des contenus. En Europe, cette pression des pouvoirs publics sur les plateformes s'est traduite par le vote de nouvelles lois : la NetzDG en Allemagne à partir de mars 2017, visant à sanctionner la diffusion de discours haineux sur les réseaux sociaux, suivie en janvier 2018 en France par la loi contre les manipulations de l'information, ciblant plus particulièrement les « *fake news* », puis en 2019 de la loi Avia, inspirée de la NetzDG, ou encore du projet législatif « *Online Harms* » au Royaume-Uni à partir d'avril 2019, qui entend limiter la violence en ligne par une régulation plus stricte des réseaux sociaux. En parallèle et en amont de ces lois, la Commission Européenne a développé à partir du printemps 2016 un code de bonne conduite relatif aux discours haineux illégaux en ligne, qui consiste en un partenariat avec les principales plateformes de réseaux sociaux présentes sur le territoire européen.

Ces différentes politiques publiques partagent des objectifs communs et incarnent à différents égards une volonté des états et des institutions européennes de reprendre la main sur le dossier ancien de la régulation des contenus sur internet (Schafer, 2018 ; Tréguer, 2019 ; Badouard, 2020). Pour autant, elles mettent en œuvre des rationalités de gouvernement et des stratégies de régulation bien différentes, relevant de la « *hard* » ou de la « *soft governance* » (Magetti, 2015), selon le degré plus ou moins contraignant des nouvelles normes qu'elles produisent. Un premier ensemble de lois peut ainsi s'apparenter à une approche de « *hard regulation* », notamment lorsque celles-ci entendent imposer aux plateformes des obligations de résultats et/ou de délais à respecter pour le retrait des contenus problématiques, et qu'elles prévoient des sanctions en cas de non mise en conformité avec les nouvelles règles. C'est notamment la voie suivie par l'Allemagne et la France (dans le cadre de la loi Avia). La seconde approche, que l'on peut qualifier de « *soft regulation* », est celle adoptée par la Commission Européenne : elle consiste en la mise en place d'un partenariat avec les plateformes définissant des obligations de moyens, assorties d'un contrôle « lâche » de leur mise en œuvre, c'est-à-dire sans sanction en cas de non mise en conformité. Le Royaume-Uni adopte quant à lui une voie intermédiaire, celle d'une obligation de moyens (et non de résultats), assorti d'un contrôle stricte, comprenant une menace d'amendes si les plateformes n'obtempèrent pas.

Cet article se propose d'appréhender l'évolution des relations entre puissances publique et technologique en Europe, à travers l'étude des réformes des politiques de modération de Facebook et YouTube sur le continent entre 2016 et 2020. Deux questions guident l'analyse

présentée ici : d'une part, comment les injonctions des pouvoirs publics en termes de régulation sont traduites par les plateformes au sein de leurs dispositifs de modération ; d'autre part, comment l'évolution de ces dispositifs est encadrée par les pouvoirs publics. Pour explorer ces deux questionnements, l'étude se focalise plus particulièrement sur les procédures de signalement et d'évaluation de deux types de contenus problématiques, à savoir les « fausses informations » et les discours de haine.

La première partie de l'article aborde la manière dont Facebook et YouTube adaptent leurs procédures de signalement à de nouvelles contraintes légales, en développant notamment des dispositifs de modération « à deux étages » pour se mettre en conformité avec les nouvelles législations européennes. La seconde partie analyse les politiques d'invisibilisation des « contenus gris » mises en œuvre par les deux firmes : pour gérer les contenus problématiques qui n'enfreignent pas directement les règlements des plateformes, celles-ci paramètrent la diffusion de ces contenus afin de limiter au maximum leur visibilité. La troisième partie traite de la question de l'automatisation de la modération, à travers le recours à l'intelligence artificielle pour détecter les discours haineux et les fausses informations. La dernière partie de l'article étudie la manière dont les pouvoirs publics encadrent ces logiques d'automatisation et d'invisibilisation par des exigences de « transparence », et revient sur les principales controverses qui entourent les différentes approches de la régulation des contenus en Europe.

L'étude proposée dans cet article repose principalement sur l'analyse de documents produits par Facebook et Google concernant leurs politiques de modération (n=76) : rapports de transparence, livres blancs, standards de publication, articles et tribunes sur les blogs officiels des firmes notamment. Cette première analyse est complétée par des entretiens réalisés avec des responsables des politiques de modération dans les deux firmes (n=3) ainsi que des observations de conférences de presse ou de tables-rondes organisées, ou auxquelles ont participé, des acteurs privés ou publics directement impliqués dans les politiques de régulation des contenus (n=11). A ce premier corpus s'ajoute un second, composé de documents produits par les autorités publiques en France, en Allemagne, au Royaume-Uni et par la Commission Européenne, portant sur la régulation des contenus sur les réseaux sociaux (n=40) : rapports gouvernementaux, textes de loi, évaluations des politiques publiques notamment.

Si les pratiques de modération sur les réseaux sociaux ont fait l'objet de travaux importants en sciences sociales ces dernières années (Roberts, 2019 ; Gillepsie 2018 ; Myers-West, 2017), la question de l'évolution des politiques des plateformes en la matière, questionnée à travers l'étude des relations entre firmes du web et pouvoirs publics, est moins documentée dans la littérature académique. Comme le note Tristan Mattelart dans son analyse des stratégies de Facebook vis-à-vis des médias d'information (Mattelart, 2020), considérer les plateformes non seulement comme des espaces où prennent place des activités sociales, mais aussi comme des acteurs politiques et économiques à part entière, nécessite de prendre au sérieux les documents qu'elles produisent pour expliciter et justifier leurs stratégies. Or, les sciences sociales se sont encore peu saisies de ces sources, peut-être parce qu'elles constituent majoritairement des supports de communication et de promotion de leurs actions pour les plateformes. Pour autant, appréhendées avec une posture critique qui les resitue dans leur contexte de production, elles demeurent des ressources précieuses pour appréhender ce pan de l'histoire politique du web qui s'écrit sous nos yeux.

Réformer les procédures de signalement

Parmi les lois et projets législatifs analysés dans cet article, la loi allemande de régulation des réseaux sociaux, surnommée NetzDG, est la plus ancienne. Entrée en vigueur le 1er janvier 2018, elle exige que les réseaux sociaux, comptant plus de deux millions d'utilisateurs sur le territoire allemand, suppriment les contenus haineux qui leur sont signalés par leurs utilisateurs dans un délai de 24 heures, sous peine d'amendes[1]. Incarnant la voie de la « *hard regulation* », la loi allemande prévoit ainsi 21 nouvelles incriminations pouvant faire l'objet d'un retrait, certaines correspondant déjà à des types de contenus prohibés par les plateformes, d'autres devant faire l'objet de nouvelles procédures (comme par exemple le recours à des symboles d' « organismes inconstitutionnels »)[2]. La NetzDG impose aux plateformes de mettre en place des dispositifs de signalement spécifiques concernant ces contenus et les oblige également à publier tous les six mois des rapports de transparence quant à la mise en application de ces nouvelles mesures[3]. Par ailleurs, le dispositif de signalement des contenus prohibés doit être aménagé vers davantage de transparence : l'internaute qui signale doit recevoir un accusé de réception ainsi qu'un nouveau message une fois la décision prise concernant le contenu signalé. L'internaute ayant publié le contenu mis en cause doit également recevoir une notification l'informant de la mise en œuvre d'une procédure d'évaluation suite à un signalement tiers.

Pour répondre à ces exigences, YouTube et Facebook s'engagent dans un processus de mise en conformité qui se traduit par la mise en place des dispositifs de modération à deux étages. Ceux-ci consistent à supprimer de la plateforme les contenus qui enfreignent à la fois les standards de publication et la loi allemande, et à bloquer géographiquement les publications qui ne contreviennent qu'aux dispositions de la NetzDG. Concrètement, dans le deuxième cas de figure, le contenu mis en cause sera retiré du Facebook allemand, mais sera toujours accessible depuis les autres déclinaisons nationales de la plateforme. Cette logique de blocage géographique n'est pas spécifique au cas allemand, et constitue davantage une stratégie globale de gestion des contenus problématiques. Dès 2010, Google assume l'idée de se plier aux exigences des gouvernements, en désindexant de son moteur de recherche et en supprimant localement sur YouTube des contenus qui contreviendraient à des lois nationales, tout en les laissant accessibles aux internautes situés dans des pays tiers[4]. Une politique similaire est mise en place sur Facebook, sans qu'il nous ait été possible d'identifier la date précise de sa mise en œuvre. Les données présentes sur la plateforme Transparence du réseau social font état de restrictions locales de contenus à partir de juillet 2013[5] et la responsable « *Global policy management* » de la firme détaille la politique de Facebook en la matière dans un document datant de mars 2015[6]. Dans les deux cas donc, le blocage géographique correspond à une démarche de mise en conformité des politiques des plateformes avec les lois nationales qui est antérieure aux cas des fausses informations et des discours haineux étudiés dans cet article.

---

[1] Loi disponible sur le site du Bundesamt für Justiz : https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html
Une version en anglais est disponible à l'adresse suivante : https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf

[2] La liste de ces nouvelles incriminations est disponible ici : https://www.facebook.com/help/285230728652028

[3] Pour Facebook, le dernier rapport est consultable à l'adresse suivante : https://about.fb.com/wp-content/uploads/2020/01/facebook_netzdg_January_2020_english.pdf
Et pour YouTube : https://transparencyreport.google.com/netzdg/youtube?hl=fr

[4] Google, « Controversial content and free expression on the web: a refresher », *Google Official Blog*, 19/04/2010.

[5] Facebook, « Content Restrictions Based on Local Law », *Facebook Transparency*, consulté le 1/09/2020.

[6] Facebook, « Explaining Our Community Standards and Approach to Government Requests », *Facebook Newsroom*, 15/03/2015.

Les mesures comprises dans la NetzDG impliquent par ailleurs, du côté des plateformes, un certain nombre d'évolutions de leurs dispositifs de modération, qui passent notamment par l'embauche et la formation de modérateurs spécialisés. Sur Facebook par exemple, un contenu signalé par un internaute comme contrevenant aux dispositions de la NetzDG fait l'objet d'une double évaluation[7] : la première, traditionnelle, réalisée par un des modérateurs de la firme au sein de la « *Community Operations Team* », qui vise à décider si le contenu contrevient aux standards de publication de la plateforme ; la seconde, dans le cas où le contenu serait conforme à ces standards, par une équipe spécialisée localisée à Dublin et à Moutain View, la « *Legal Takedown Request Operations Team* », afin de statuer sur la décision à prendre concernant le contenu en question. Tout l'enjeu pour Facebook est de respecter le délai de 24h pour l'évaluation des signalements imposé par la NetzDG. Pour ce faire, la firme a annoncé au 30 juin 2019[8] avoir formé 80 modérateurs aux dispositions spécifiques de la loi allemande. Le même mécanisme de double évaluation a été mis en place par Google sur YouTube[9]. Les signalements spécifiques à la NetzDG sont ainsi étudiés par une équipe de 73 membres formés à l'évaluation des infractions pénales relevant spécifiquement de cette loi, et résidant en Allemagne.

Une autre évolution du dispositif de modération portée par la NetzDG a trait aux applications permettant à un internaute de signaler un contenu indésirable. Ici, les politiques de Facebook et Google diffèrent. Sur Facebook, signaler un contenu au nom de la NetzDG implique de passer par un formulaire de plainte spécifique, qui n'est pas directement accessible depuis le fil d'actualité, et qui n'est disponible qu'en Allemagne[10]. Sur YouTube à l'inverse, les nouvelles fonctionnalités de signalement sont intégrées directement au dispositif de visionnage : un item représentant trois petits points, situé sous chaque vidéo, permet de faire défiler un menu listant un ensemble d'actions à réaliser, parmi lesquelles le signalement de la vidéo en question. Lorsqu'un contenu est signalé, il est demandé à l'internaute de qualifier le contenu qu'il dénonce (« contenu à caractère sexuel », « contenu offensant ou haineux », etc.) : en Allemagne, il est proposé aux internautes de signaler ce contenu au nom de la NetzDG. Pour simplifier le dispositif, les équipes de YouTube ont ramené les 21 nouvelles incriminations à un ensemble de 7 catégories différentes. Le fait d'intégrer cette possibilité directement dans le design du dispositif de signalement a une conséquence directe : celle d'augmenter fortement le nombre des requêtes effectuées par les internautes. Alors que dans son premier rapport d'application[11], Facebook précisait avoir reçu quelques centaines de signalements au nom de la NetzDG, Google annonçait dans le même temps en avoir reçu des centaines de milliers[12], concernant dans la plupart des cas des incitations à la haine ou de l'extrémisme politique.

La loi française sur les manipulations de l'information, entrée en vigueur le 22 décembre 2018[13], n'implique pas une telle évolution des dispositifs de signalement des plateformes,

---

[7] Facebook, *NetzDG Transparency Report*, Juillet 2018.
[8] Facebook, Ibid.
[9] Google, « Suppression de contenu en vertu de la loi NetzDG », *Google Transparence des informations*, consultée le 01/09/2020.
[10] Le NetzDg help center est en revanche accessible depuis tous les pays : https://www.facebook.com/help/285230728652028
[11] Facebook, *NetzDG Transparency Report*, Juillet 2018.
[12] Google, « Suppression de contenu en vertu de la loi NetzDG », *Google Transparence des informations*, consultée le 01/09/2020.
[13] Legifrance, LOI n° 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information, consulté le 01/09/2020.

dans la mesure où son objectif principal réside dans la création d'une procédure judiciaire extérieure aux dispositifs de modération. Cette procédure doit permettre, en période électorale, la saisie d'un juge des référés qui peut exiger le retrait d'une information relevant « d'allégations ou imputations inexactes ou trompeuses d'un fait de nature à altérer la sincérité du scrutin à venir (et qui sont) diffusées de manière délibérée, artificielle ou automatisée et massive par le biais d'un service de communication au public en ligne ». Le juge saisi dispose alors d'un délai de 48h pour statuer sur le retrait d'une information mise en cause. En dehors des périodes électorales, les opérateurs sont tenus à un « devoir de coopération » qui s'exprime notamment par la mise en place, par les plateformes, d'un dispositif de signalement spécifique aux fausses informations. Sur Facebook par exemple, la catégorie « fausse information » a été ajoutée aux motifs de signalement des contenus, accessible depuis chaque publication sur les fils d'actualité des internautes. Sur YouTube, la mention « contenu trompeur » a été associée au motif « spam », sans que nous puissions assurer que cette mention est liée à l'application de la loi française sur les manipulations de l'information.

En France, la manière dont les plateformes doivent rendre compte du respect de leur « devoir de coopération » diffère du cas allemand. La loi sur les manipulations de l'information stipule que le contrôle des mesures mises en œuvre par les plateformes sur le territoire national incombe au Conseil Supérieur de l'Audiovisuel (CSA), qui doit régulièrement publier des rapports sur les modalités de mise en œuvre de la loi. En Allemagne, les plateformes s'acquittent directement de la publication de ces rapports, qui sont contrôlés par l'institution judiciaire. Dans le cadre de ses nouvelles fonctions, le CSA a publié le 15 mai 2019 une première recommandation à destination des plateformes visant à leur mise en conformité avec la nouvelle loi[14]. Cette première recommandation a été doublée le 27 février 2020 d'un questionnaire précisant les données auxquelles le Conseil souhaite accéder[15]. Parmi ces informations, on retrouve notamment la présentation « précise et ergonomique » du dispositif de signalement, des statistiques sur le nombre de requêtes effectuées par les internautes, la durée de leur traitement, les recours offerts aux internautes ou encore les budgets alloués à ces différentes tâches.

La loi Avia en revanche, qui ciblait les discours de haine, impliquait une réforme plus stricte des procédures de signalement[16]. D'abord, elle entendait imposer aux entreprises du web la mise en place d'un bouton de notification unique pour toutes les plateformes. Comme pour la NetzDG, dont elle s'inspirait largement, elle visait également à renforcer la transparence de la procédure : accusé de réception pour l'internaute qui signale, notification du signalement pour l'internaute mis en cause, information des deux parties une fois la décision prise. Contrairement à la NetzDG, la loi Avia impliquait également la mise en place d'une procédure d'appel pour les internautes, leur permettant de demander une seconde évaluation des publications supprimées. L'application de ces mesures n'est plus à l'ordre du jour au moment où cet article est écrit, celles-ci ayant été censurées par le Conseil Constitutionnel le

[14] Legifrance, Recommandation n° 2019-03 du 15 mai 2019 du Conseil supérieur de l'audiovisuel aux opérateurs de plateforme en ligne dans le cadre du devoir de coopération en matière de lutte contre la diffusion de fausses informations, consulté le 01/09/2020.
[15] CSA, *Lutte contre la manipulation de l'information : questionnaire aux opérateurs de plateformes en ligne*, consulté le 01/09/2020.
[16] Assemblée nationale, *Proposition de loi visant à lutter contre les contenus haineux sur internet*, version adoptée par l'Assemblée nationale en seconde lecture, 22 janvier 2020.

18 juin 2020[17]. Une telle procédure d'appel existe cependant déjà sur YouTube depuis juillet 2010[18], et sur Facebook depuis avril 2018[19].

La loi britannique « Online Harms » prévoit également une réforme des procédures de signalement et d'appel des plateformes, en leur imposant de nouveaux standards en la matière (notamment en termes de typologie des contenus signalés, de transparence de la procédure et de rapidité de traitement). Peu de détails sur cette réforme spécifique sont donnés par les autorités britanniques à cette étape, le processus législatif en étant encore à un stade précoce, par ailleurs ralenti par la crise sanitaire de 2020 : après la publication d'un livre blanc en avril 2019[20], une consultation publique a été menée d'avril à juillet, suivie d'une publication d'une synthèse des contributions le 12 février 2020[21]. La nouvelle loi, attendue initialement pour le printemps 2020, a été repoussée à la fin de l'année 2021[22]. Le code de conduite de l'Union Européenne, qui repose également sur une logique d'intensification des traitements des signalements[23], implique par ailleurs du côté des signataires le respect d'un certain nombre de standards minimaux concernant les procédures en question, comme la publication de règlements de la communauté, la transparence des décisions prises et l'information des internautes concernés. Nous reviendrons en détail sur le fonctionnement de ce code plus tard dans l'article.

A ce stade, une première conclusion peut être tirée de l'analyse de l'évolution des procédures de signalement. Les pressions des états et des institutions européennes, en particulier les approches relevant de la « *hard regulation* », produisent un certain nombre d'effets, qui se traduisent à la fois par une évolution du design des dispositifs (boutons ou formulaires spécifiques), par la création de nouvelles catégories de propos prohibés et par la formation de personnels spécialisés dans leur modération. Pour se mettre en conformité avec ces exigences, les plateformes créent des déclinaisons nationales de leurs services, qui diffèrent des autres déclinaisons existantes. Cet « emboitement géographique » des plateformes illustre à la fois la difficulté qu'il y a, du côté des firmes, à fixer des règles de publication valables à une échelle internationale (Gillepsie, 2018), et démontre également la capacité des états à fixer des règles d'usage spécifiques à l'intérieur de leurs propres frontières.

Mesures préconisées ou mises en œuvre concernant les procédures de signalement

Loi	Pays concerné	Mesures préconisées / mises en vigueur	Traduction par les plateformes	Etat d'avancement de loi
NetzDG	Allemagne	-Création de 21 nouvelles incriminations pouvant faire l'objet d'un retrait -Réforme des procédures de signalement (délais de retrait notamment)	-Nouvelles catégories de signalement -Embauche et formation de personnels qualifiés -Adaptation du dispositif de signalement aux nouvelles	Entrée en vigueur le 1er janvier 2018

[17] Conseil Constitutionnel, Décision n°2020-801 DC du 18 juin 2020 relative à la Loi visant à lutter contre les contenus haineux sur internet.
[18] Google, « Strike you're out! Or maybe not? », *YouTube Official Blog*, 02/07/2010.
[19] Facebook, « Standards de la Communauté : Facebook publie ses directives internes et élargit son processus de recours aux contenus individuels », *Facebook Newsroom*, 24/04/2018.
[20] Gouvernement du Royaume-Uni, *Online Harm White Paper*, avril 2019.
[21] Gouvernement du Royaume-Uni, *Online Harm White Paper - Initial consultation response*, février 2020.
[22] BBC News, "Online Harms bill :Warning Over "unacceptable" delay", 29 juin 2020, consulté le 01/09/2020 à l'adresse : https://www.bbc.com/news/technology-53222665
[23] Commission Européenne, *Code of Conduit on countering illegal hate speech online*, 30 juin 2016.

		-Publication de rapports de transparence tous les six mois -Amendes en cas de non mise en conformité	exigences -Publication en ligne de rapports de transparence	
Loi sur les manipulations de l'information	France	-Création d'un nouveau délit de propagation de fausses nouvelles sur Internet -Création d'une procédure judiciaire spécifique à la modération des fausses nouvelles sur internet -Devoir de coopération des plateformes	-Nouvelle catégorie de signalement -Adaptation du dispositif de signalement -Publication de rapports de transparence auprès du CSA	Entrée en vigueur le 22 décembre 2018
Loi de lutte contre les discours de haine (Loi Avia)	France	-Création d'un bouton unique de signalement commun à l'ensemble des plateformes -Réforme des procédures de signalement et d'appel (délais de retrait notamment) -Devoir de coopération des plateformes -Amendes en cas de non mise en conformité	/	Projet de loi censuré par le Conseil Constitutionnel en juin 2020.
Online Harms White Paper	Royaume-Uni	-Duty of Care / obligation de moyens -Mise en conformité des procédures de signalement -Devoir de coopération -Amendes en cas de non mise en conformité	/	Loi en cours d'élaboration
Code de conduite pour lutter contre la haine en ligne	Union Européenne	-Mise en place de standards de publication et de procédures de signalement -Respect des lois nationales -Respect d'un délai maximum de 24h pour le retrait des contenus haineux -Partenariats avec des associations (« trusted reporters ») -Mise à jour des formations des modérateurs -Méthode de testing	-Augmentation de la part des signalements traités -Diminution des délais de traitement -Partenariats avec des associations (« trusted flaggers »)	Partenariat mis en place en mai 2016

L'invisibilisation des contenus problématiques

Le rapport des entreprises du numérique à ces nouvelles normes légales n'est ni univoque, ni linéaire, et les motivations qui les poussent à se conformer à ces normes sont variées. D'une part, les injonctions étatiques peuvent rencontrer des intérêts économiques qui vont pousser

les plateformes à collaborer. Comme nous le confiait l'un des responsables des politiques publiques chez Google à propos de la législation française sur les « *fake news* » : « Quand on regarde Google et YouTube, leur première mission, en termes d'*imperative business*, c'est de fournir aux gens des informations qui leur soient utiles. Il y a une menace concrète pour le business s'il y a une impression générale qu'on est pas capable de faire ça, et que l'information que nous fournissons n'est pas fiable »[24]. D'autre part, si les plateformes adoptent globalement une attitude coopérative par rapport aux nouvelles législations, cette attitude se concrétise à travers la mise en œuvre de différentes stratégies : certaines entreprises peuvent anticiper les projets de loi et agir en amont des discussions parlementaires, à partir des rapports publiés par le gouvernement ou des annonces de personnalités politiques, adopter une logique de mise en conformité après le vote des lois, en se pliant aux nouvelles mesures, ou encore accompagner la réalisation de la loi pendant les travaux des Commissions, dans un processus itératif, lors d'auditions, d'activités de lobbying ou dans des formes de collaboration plus originales comme celles mises en œuvre par la mission Loutrel en France[25]. Notre même contact chez Google exprimait la situation en ces mots : « La question n'est pas de savoir si les entreprises ont des motivations internes ou si la loi a un réel pouvoir de contrainte : c'est souvent les deux. Les lois ont des conséquences très concrètes sur notre travail, mais on n'attend pas nécessairement la loi, et parfois on va au-delà de la loi. C'est un mélange des deux, et c'est toujours difficile de dire ce qui est lié à une motivation strictement interne ou strictement externe ».

Le cas des discours de haine est représentatif de cette articulation entre motivations internes et pressions externes sur les dispositifs de modération : depuis la création de YouTube et de Facebook au milieu des années 2000, les propos haineux font partie des types de discours prohibés. Pour autant, les controverses publiques sur la violence et le racisme en ligne, les procédures judiciaires mises en place contre les grandes firmes du web aux Etats-Unis, ou les projets de régulation portés par les pays européens, produisent contraintes et pressions qui, selon nos interviewés, les amènent à réformer leurs dispositifs de modération et à faire évoluer la manière dont ils communiquent à leurs sujets. Pour Facebook comme pour Google, un tournant semble ainsi s'opérer en avril 2018. Le 24 avril, Facebook publie en effet pour la première fois ses « standards de la communauté », soit un ensemble de documents autrefois réservés aux modérateurs sur l'application des directives de publication[26], et organise des conférences publiques visant à expliquer comment les contenus sont modérés sur la plateforme. Dans ce contexte, l'entreprise précise quels types de propos sont considérés comme des discours de haine (principalement l'appel à la violence, la déshumanisation, l'infériorisation, l'incitation aux discriminations, les insultes envers quelqu'un en raison de son origine, sa religion, son orientation sexuelle, son sexe, sa caste ou sa situation de handicap). Lors des conférences publiques, des tests participatifs sont même réalisés avec le public pour mettre en avant les difficultés que peuvent rencontrer les modérateurs sur le terrain[27].

La veille de la publication des standards de publication de Facebook, le 23 avril, Google publie pour la première fois un rapport sur « l'application du règlement de la communauté

[24] Entretien réalisé le 6 juillet 2020.
[25] La mission Loutrel consiste en un partenariat inédit entre le gouvernement français et Facebook pour lutter contre la haine en ligne, sur lequel nous revenons plus tard dans l'article.
[26] Facebook, « Standards de la Communauté : Facebook publie ses directives internes et élargit son processus de recours aux contenus individuels », *Facebook Newsroom*, 24/04/2018.
[27] Comme cela a été le cas lors de la présentation publique des standards de publication dans les locaux de Facebook à Paris le 15/05/2018.

YouTube »[28] , qui explicite et quantifie le travail réalisé par les modérateurs en fonction, notamment, des types de propos visés. Facebook suivra le mois suivant en publiant également son premier rapport sur l'application de ses standards de publication. A partir de cette date, les deux entreprises mettront à jour tous les trois mois, via des plateformes interactives, des rapports de transparence sur l'application de leurs politiques de modération. Ces efforts de transparence poussent à mettre en avant le durcissement de ces politiques et la fermeté de leur application. Concernant les propos haineux par exemple, Facebook annonce dans ses rapports qu'au dernier semestre 2017, 1,6 millions de contenus ont fait l'objet d'une évaluation pour « *hate speech* », contre 22,5 millions au deuxième trimestre 2020. Entre les deux dates, l'augmentation du nombre d'évaluations est continue[29]. Google annonce de son côté 18 950 vidéos supprimées et 1713 chaînes clôturées sur YouTube au dernier trimestre 2018 pour contenus haineux, contre 80 033 vidéos et 3308 chaînes au second trimestre 2020[30]. Cette intensification de la modération y concerne également les commentaires : plus de 30 millions de commentaires haineux sont retirés au 3ème trimestre 2019 contre plus de 150 millions au second trimestre 2020.

Les discours de haine constituent des contenus relativement faciles à modérer lorsqu'ils sont explicites. Il en va autrement lorsque ceux-ci sont masqués, notamment via des techniques d'offuscation (fautes d'orthographes volontaires, noms de code, sous-entendus) (Gillepsie, 2018 ; Roberts, 2019). Les propos tenus entrent alors dans une zone grise : ils ne contreviennent pas explicitement aux standards de publication des plateformes, mais constituent pour autant des contenus indésirables. Cette ambiguïté est encore plus forte du côté des « *fake news* » : une fausse information, à moins d'être insultante ou diffamatoire, n'enfreint pas les règles de publication mais pour autant, les plateformes, pour des raisons politiques ou économiques, cherchent à limiter leur propagation. Pour gérer ces « contenus gris », les deux plateformes ont mis en place des stratégies d'invisibilisation, qui visent à limiter leur diffusion sans pour autant les supprimer.

Chez Facebook, la stratégie en question se résume en trois mots : « *remove, reduce, inform* ». Si la firme annonce l'avoir mise en place à partir de 2016, la politique est assumée, explicitée et renforcée à partir d'avril 2019[31]. Son principe est de retirer de la plateforme les contenus qui contreviennent aux standards de publication ainsi qu'aux lois nationales en vigueur dans certains pays (*remove*), d'informer les usagers sur les contenus qu'ils consultent et leur degré de fiabilité (*inform*), et de « dégrader la visibilité » des contenus gris, identifiés comme étant de mauvaise qualité, mais n'étant ni illégaux ni ne contrevenant aux standards de publication (*reduce*). Concrètement, si ces contenus sont toujours en ligne sur les plateformes, ils s'affichent plus bas dans les fils d'actualité, et sont donc moins vus et moins partagés par les internautes. En septembre 2019, Google introduit de son côté sa politique dite des « 4Rs », pour « *Remove, Raise, Reward and Reduce* » sur YouTube[32]. Son principe est de retirer de la plateforme les contenus qui violent les standards (*remove*), d'offrir des espaces de visibilité particulier aux sources fiables (*raise*), de donner des bonus de visibilité aux créateurs de contenus fiables (*reward*), tout en limitant la visibilité des contenus peu fiables et mal évalués (*reduce*). Sur YouTube, cette dégradation de la visibilité s'exprime différemment : les

[28] Facebook, « Facebook Publishes Enforcement Numbers for the First Time », *Facebook Newsroom*, 15/05/2018.
[29] Facebook, *Community Standards Enforcement Report*, consulté le 01/09/2020.
[30] Google, *Application du règlement de la communauté YouTube*, consulté le 01/09/2020
[31] Facebook, « Remove, Reduce, Inform : New Steps to Manage Problematic Content », *Facebook Newsroom*, 10/04/2019.
[32] Google, « The Four Rs of Responsability, Part 1: Removing harmful content », *YouTube Official Blog*, 03/09/2019.

contenus problématiques sont moins bien classés par les algorithmes de recommandation, voire sont rendus impossible à partager[33]. Le résultat est cependant similaire à Facebook : les vidéos sont moins vues et moins partagées, tout en restant en ligne sur la plateforme. D'après Google, cette technique permettrait de réduire de 80% le visionnage de vidéos problématiques[34].

Cette gestion des contenus gris pose un problème majeur du côté des usagers : en cas d'erreur de modération, les décisions des plateformes sont au final peu visibles et difficilement contestables (puisque l'internaute continue de publier, mais ne touche pas son public habituel). Pour faire face aux risques d'accusation de censure abusive, Google comme Facebook adoptent une logique similaire : déléguer à des acteurs tiers, indépendants des plateformes, la responsabilité de statuer sur la nocivité de ces contenus gris.

Concernant les fausses informations, Facebook a ainsi mis en place un partenariat avec des journalistes professionnels, le « *Third-Party Fact-checking program*»[35], lancé en décembre 2016, et qui comprenait trois ans plus tard 50 partenaires couvrant 40 langues différentes. Le principe de ce partenariat est d'embaucher au sein de rédactions reconnues des journalistes membres de l'International Fact-Checking Network, qui travaillent directement pour la plateforme. Concrètement, sur Facebook, des mécanismes de reconnaissance automatique permettent d'identifier des signaux faibles indiquant le manque de fiabilité d'une information (le signalement des utilisateurs par exemple, ou des commentaires dans lesquels les internautes en contestent la véracité). Lorsque ces signaux faibles sont activés, la publication mise en cause est transmise à l'équipe de journalistes fact-checkers rémunérée par la plateforme, qui doivent la vérifier et lui attribuer une note de fiabilité[36]. Si la note en question est mauvaise, trois décisions peuvent être prises concernant la publication[37] : réduire sa visibilité en faisant en sorte que l'information s'affiche plus bas dans les fils d'actualité ; produire un « avertissement de partage » qui s'affiche lorsqu'un internaute souhaite partager l'information en question afin de l'avertir de son caractère douteux ; appliquer un « avertissement de consultation » directement sur le contenu et qui nécessite un second clic pour pouvoir y accéder. D'après des chiffres livrés par une des responsables des politiques publiques de Facebook France lors de notre entretien[38], 50 millions d'avertissements auraient été associés à des contenus à l'échelle internationale en avril 2020, qui auraient dissuadé 95% des utilisateurs de cliquer sur les contenus en question. Nous n'avons pas eu les moyens de croiser ces informations ou de les vérifier par nous-mêmes.

Chez Google, concernant le moteur de recherche développé par la firme, ce ne sont pas des journalistes, mais des internautes ordinaires qui se voient attribuer la fonction d'évaluer collectivement la fiabilité des contenus. Ces internautes sont réunis au sein d'un panel, les « *quality raters* »[39], qui évaluent leur expérience de navigation et de consommation de contenus lorsqu'ils utilisent les produits de la firme. Entre autres choses, on leur demande par exemple de se prononcer sur le caractère intrusif des publicités sur les sites, le degré

[33] Google, « Our ongoing work to tackle hate », *YouTube Official Blog*, 05/06/2019.
[34] Ibid.
[35] Facebook, « Programme Facebook de vérification des informations par des tiers », *Facebook Journalism Project*, consulté le 01/09/2020.
[36] Facebook, « Hard Questions: How is Facebook's Fact-Checking Program Working? », *Facebook Newsroom*, 14/06/2018.
[37] Facebook, « Fonctionnement de notre programme de vérification des informations », *Facebook Journalism Project*, 11/08/2020.
[38] Entretien réalisé le 17 juillet 2020.
[39] Google, « Our latest quality improvements for Search », *Google Blog The Keyword*, 25/04/2017.

d'expertise de la page qu'ils consultent ou encore l'adéquation entre ce qu'ils lisent et ce qu'ils ont recherché. Si Google prétend que les évaluations des *quality raters* n'ont pas d'impact direct sur le référencement d'un site en particulier, la firme avoue également que leurs résultats ont des effets significatifs sur les mises à jour de l'algorithme de classement, afin de dégrader la visibilité des contenus identifiés comme étant de mauvaise qualité[40]. La même logique d'évaluation par des internautes régit par ailleurs les différentes mises à jour de l'algorithme de recommandation de YouTube. Régulièrement mis en cause pour favoriser des contenus complotistes (Faddoul et al., 2020), l'algorithme a subi une modification majeure à partir de janvier 2019, visant à associer au temps de visionnage des vidéos, des données relatives aux commentaires et aux réactions des internautes, ainsi que les évaluations manuelles des *quality raters*, afin d'améliorer la qualité de ces recommandations[41]. D'après une étude d'une équipe de chercheurs de Berkeley, non encore publiée (Faddoul et al., 2020), cette modification aurait permis de réduire de moitié la recommandation de contenus jugés complotistes.

L'invisibilisation des contenus gris repose donc sur des formes de régulation par le design des outils de consultation. L'enjeu est ici d'influencer les comportements des internautes, soit explicitement via des avertissements les dissuadant de consulter ou partager une information, soit implicitement via les algorithmes de classement des informations qui font en sorte que ces derniers soient moins vus (en dégradant leur visibilité dans les fils d'actualité sur Facebook, en les recommandant moins sur YouTube). La mise en (in-)visibilité des informations devient ainsi un outil de régulation : il s'agit moins d'empêcher qu'un contenu soit publié que de le couper d'une audience potentielle. Cette évolution est moins anodine qu'il n'y paraît, dans la mesure où le tri et le classement des informations est au cœur même des débats sur la responsabilité éditoriale des plateformes (Grimmelmann, 2014 ; Sire, 2015).

En France, par exemple, la loi de confiance dans l'économie numérique, votée en 2004, instaure une distinction entre fournisseurs d'accès (considérés comme ayant la responsabilité pénale la moins importante), hébergeurs (qui disposent aux yeux de la loi d'une « responsabilité atténuée »), et éditeurs (considérés comme pleinement responsables). L'application de cette législation aux activités des plateformes a mené à des décisions de justice contradictoires au cours des années 2000 (Badouard, 2020) : dans certaines affaires, les plateformes étaient considérées comme des hébergeurs, car n'exerçant pas de contrôle a priori sur les contenus publiés via leurs outils, et dans d'autres cas comme des éditeurs, leurs algorithmes remplissant des fonctions de filtrage et de tri pouvant s'apparenter à des activités éditoriales. La loi République numérique, votée en 2015, entendait résoudre ce problème à travers l'instauration du statut d' « opérateur de plateforme ». Aujourd'hui, le recours aux ressources et infrastructures de la visibilité pour exercer des fonctions non plus uniquement de curation, mais de modération, conforte le rôle typiquement éditorial des firmes, qui via des acteurs externes dont elles s'assurent la légitimité, décident indirectement de ce qui mérite d'être vu ou oublié, assumant de fait un rôle de *gatekeeper* sur les réseaux sociaux (Barzilai-Nahon, 2008).

L'automatisation de la modération

[40] Google, *Search Quality Evaluator Guidelines*, 05/12/2019.
[41] Google, "Continuing our work to improve recommendations on YouTube", You Tube Official Blog, 25/01/2019.

Pour identifier les contenus problématiques à modérer, les plateformes ont historiquement recours aux signalements des internautes. Face à certains types de publications qui nécessitent des formes de modération *en contexte*, comme les fausses nouvelles ou les discours haineux, l'interprétation collective et dispersée des internautes présente une certaine efficacité. Pour autant, face à l'énorme quantité de contenus publiés quotidiennement sur ces plateformes, les seuls signalements ne suffisent pas. Pour faire face aux enjeux du « passage à l'échelle » de leurs dispositifs de modération, les plateformes parient sur l'intelligence artificielle et mettent en œuvre des outils de détection automatique des contenus (Gillepsie, 2018). Cette tendance est aujourd'hui confortée et encouragée par les lois qui, comme la NetzDG et le projet de loi Avia en France, ou le code de conduite de la Commission Européenne, imposent ou encouragent les plateformes à respecter un délai de 24h pour traiter les contenus qui leur sont signalés. Avoir recours à des outils qui permettent de détecter et d'éliminer un certain nombre de publications avec même qu'elles soient signalées constitue ainsi un moyen efficace de limiter la charge de modération.

Chez Facebook, la première évocation publique de ces techniques de détection automatique date de juin 2017, dans le cadre des politiques de la firme de lutte contre le terrorisme[42]. Parmi les outils utilisés, on retrouve en bonne place la reconnaissance automatique d'images et de textes : des algorithmes sont ainsi « entrainés » à reconnaître des publications problématiques, en se « nourrissant » de grandes bases de données comprenant des contenus déjà sanctionnés par les modérateurs. En septembre 2019, Facebook annonce élargir l'utilisation de ces algorithmes de détection aux contenus haineux et aux groupes racistes[43]. L'usage de la détection automatique semble se généraliser au sein de l'entreprise comme méthode principale pour lutter contre la haine en ligne, et les mêmes procédés sont appliqués à Instagram[44]. Les « *Community standards enforcement reports* » de la firme[45] confirment ainsi que la part des propos haineux détectés automatiquement (avant même qu'un utilisateur ne les signale) est passé de 23,6% au dernier trimestre 2017 à 94,5% au second trimestre 2020. Google semble suivre une voie similaire quant à la détection des vidéos et des commentaires problématiques, même si les données présentes dans les rapports d'application du règlement de la communauté YouTube[46] sont beaucoup moins précises que celles de Facebook. La firme annonce cependant que le recours à la détection automatique (*automated flagging*), qui avait déjà permis de supprimer plus de 6 millions de vidéos au dernier trimestre 2017, avait conduit à la suppression de plus de 10 millions au second trimestre 2020, sans que l'on puisse savoir spécifiquement quelle part de ces vidéos a été supprimée en raison de leur contenu haineux ou mensonger. Dans un post de décembre 2017 sur le blog officiel de YouTube, la CEO de la plateforme, Susan Wojcicki, annonçait que 98% des contenus détectés pour « extrêmisme violent » dans le dernier trimestre 2017 l'avaient été par des outils automatisés[47].

De l'aveu même des deux firmes[48], le recours à la détection automatique de contenus sensibles comme les discours de haine pose deux problèmes principaux : d'une part, son efficacité, dans la mesure où les réseaux et individus racistes, antisémites, homophobes,

[42] Facebook, « Hard Questions: How We Counter Terrorism », *Facebook Newsroom*, 15/06/2017.
[43] Facebook, « Combating Hate and Extremism », *Facebook Newsroom*, 17/09/2017.
[44] Facebook, « An Update on Combating Hate and Dangerous Organizations», *Facebook Newsroom*, 12/05/2020.
[45] Facebook, *Community Standards Enforcement Report*, consulté le 01/09/2020.
[46] Google, *Application du règlement de la communauté YouTube*, consulté le 01/09/2020
[47] Google, « Expanding our work against abuse of our platform », *YouTube Official Blog*, 05/12/2017.
[48] Facebook, « AI advances to better detect hate speech », *Facebook AI*, 12/05/2020 ; Google, « Featured Policies: Hate Speech », *Google Transparency Report*, consulté le 01/09/2020.

misogynes, savent manier les arts de la dissimulation pour échapper à la détection ; d'autre part, la détection automatique conduit à un certain nombre d'erreurs de jugement (par exemple quand un internaute condamne des propos racistes en les reproduisant). Le problème réside, pour les contenus haineux comme pour les « *fake news* », dans la prise en compte du contexte de l'échange, qui reste encore problématique dans le domaine de l'intelligence artificielle (Roberts, 2019 ; Gillepsie, 2018). Si les algorithmes peuvent se montrer extrêmement performants pour analyser les phénomènes exprimables sous la forme de données numériques, le contexte culturel et conjoncturel d'un échange ne l'est pas forcément.

Pour pallier les risques de censure abusive que fait courir l'automatisation de la modération, les plateformes valorisent aujourd'hui leurs procédures d'appel, qui permettent aux internautes de demander une seconde évaluation de leurs publications quand ils jugent que leur suppression est illégitime. Chez Facebook, la mise en place d'une procédure d'appel coïncide directement avec la publication de nouveaux standards de publication en avril 2018[49]. Cette seconde évaluation doit avoir lieu sous 24h, et la plateforme s'engage à republier le contenu si le retrait est finalement jugé abusif. Les rapports de transparence sur l'application du règlement de la communauté comporte des chiffres sur les procédures d'appel depuis le premier semestre 2019[50]. Sur l'ensemble de l'année 2019 par exemple, sur les 21,2 millions de contenus qui ont fait l'objet d'une procédure de modération pour « *hate speech* », 4,7 millions ont fait l'objet d'une procédure d'appel de la part des internautes mis en cause, soit 22,2%. Sur ces 4,7 millions d'appels, 469 700 contenus furent restaurés, soit près de 10%.

Sur YouTube, la procédure d'appel est plus ancienne et date de juillet 2010[51]. Dans ces rapports d'application du règlement de la communauté YouTube, Google livre moins de détails que Facebook sur la procédure et ces résultats (le nombre d'appels et de restaurations n'est par exemple pas ventilé par type de contenus), et les chiffres ne sont disponibles qu'à partir du dernier trimestre 2019. Ceci étant, à l'échelle de la plateforme, et tout contenu confondu, 106 587 vidéos supprimées ont fait l'objet d'un appel sur cette période, et 20 868 ont été restaurées, soit 19,6%. Sur les deux premiers semestres de 2020, ce chiffre augmente même à 41% (201 680 restaurations pour 491 380 appels), sans que l'on puisse savoir ce qui explique cette augmentation. Dans les deux cas, il est à noter que le nombre de restaurations est important, puisqu'il oscille entre 10 et 41% des contenus supprimés. Ces chiffres mettent en avant deux phénomènes : d'une part, les procédures de modération, qu'elles soient humaines ou automatisées, sont faillibles, et chaque jour, de nombreux contenus légitimes sont retirés ou invisiblisés par erreur ; d'autre part, une procédure d'appel est nécessaire pour corriger ces formes de censure abusive et permettre aux internautes de faire valoir pleinement leur droit à l'expression. Pour autant, les procédures d'appel ne sont pas une panacée. Facebook a par exemple fait mener des audits externes de ses politiques de modération par des organisations et experts de la société civile à partir de mai 2018[52]. Dans le second rapport d'audit[53], en date du 30 juin 2019, l'équipe en charge de la rédaction du rapport précisait ainsi que la procédure d'appel constituait l'un des leviers de progression de la firme : les jugements rendus n'étaient pas assez explicités, et la procédures dans son ensemble manquait globalement de transparence.

[49] Google, « Featured Policies: Hate Speech », *Google Transparency Report*, consulté le 01/09/2020.
[50] Facebook, *Community Standards Enforcement Report*, consulté le 01/09/2020.
[51] Google, « Strike you're out! Or maybe not? », *YouTube Official Blog*, 02/07/2010 ; Google, « Appeal Community Guidelines actions », *YouTube Help*, consulté le 01/09/2020.
[52] Facebook, *Update on Facebook's Civil Rights Audit*, décembre 2018.
[53] Facebook, *Facebook's Civil Rights Audit - Progress Report*, juin 2019.

Typologie des formes de modération sur YouTube et Facebook

	Acteurs / ressources	Logique de régulation interne	Formes de contrôle étatique (régulation externe)
Signalement	Internautes et modérateurs professionnels	Vigilantisme (exercice collectif de la modération)	-Information des internautes mis en cause -Transparence de la procédure de signalement -Procédures d'appel
Détection automatique	Intelligence artificielle et modérateurs professionnels	Automatisation de la modération	-Transparence des résultats -Procédures d'appel
Invisibilisation	Algorithmes de classement et de recommandation	Gouvernance par les infrastructures	-Transparence du fonctionnement des algorithmes

Face aux dynamiques d'automatisation de la modération et d'invisibilisation des contenus, qui reposent sur une régulation *par* les ressources techniques, la réponse des états réside principalement dans une injonction à la transparence. Que ce soit dans la NetzDG en Allemagne, dans la loi sur les manipulations de l'information ou sur les discours de haine en France, dans le projet « *Online Harms* » au Royaume-Uni ou dans l'application du code de bonne conduite de la Commission Européenne, les plateformes sont invitées à publier des rapports explicitant le fonctionnement de leurs algorithmes de classement, la manière dont ceux-ci influencent la consommation d'information par les internautes, les moyens techniques mis en œuvre pour détecter les contenus problématiques ou encore le fonctionnement de leurs régies publicitaires. Pour autant, les moyens mis en œuvre pour contrôler ces rapports de transparence varient grandement d'un pays à l'autre.

Les deux lois françaises stipulent, comme nous l'avons vu, que la mise en conformité des plateformes vis-à-vis des nouvelles mesures doit être contrôlée par le Conseil Supérieur de l'Audiovisuel. Au-delà des informations relatives aux dispositifs de signalement des plateformes que nous avons évoquées plus haut, les rapports exigés par le CSA comprennent un ensemble d'exigences liées à la transparence des algorithmes de classement et à leurs modes de fonctionnement, aux données utilisées par les algorithmes de recommandation ou encore aux différents dispositifs d'information des usagers concernant ces classements et recommandations[54]. La loi Avia, dans sa dernière mouture votée à l'Assemblée Nationale, impliquait également que le CSA puisse accéder « aux principes et méthodes de conception des algorithmes ainsi qu'aux données utilisées par ces algorithmes pour se conformer à ces nouvelles obligations », sans que la méthode utilisée pour ce faire ne soit précisée. Par ailleurs, dans le cadre de la loi contre les « *fake news* », le CSA ne dispose pas de pouvoir de contrainte sur les plateformes, et aucune sanction n'est prévue dans le texte si les opérateurs n'agissent pas conformément à ce qui leur est demandé. La loi Avia, en revanche, prévoyait que le CSA puisse infliger des amendes aux plateformes si celles-ci ne respectaient pas la

[54] CSA, *Lutte contre la manipulation de l'information : questionnaire aux opérateurs de plateformes en ligne*, consulté le 01/09/2020.

nouvelle législation. Dans sa version censurée par le Conseil Constitutionnel publié au Journal Officiel, le seul moyen d'action restant au CSA est la mise en place d'un Observatoire de la haine en ligne, ne disposant d'aucun pouvoir de contrainte sur les activités des plateformes.

En Allemagne, les plateformes doivent produire tous les six mois des rapports sur leurs activités en matière de modération, qui comportent des données relatives aux seuls signalements et à leur traitement. Le contrôle publique des activités des plateformes est supervisé par le ministère de la justice, via des organisations et des institutions qu'il mandate. En cas de non-respect des mesures, ou de transmission de données falsifiées, les plateformes peuvent se voir infliger de lourdes amendes- sur lesquelles nous reviendrons, sans que le texte de loi stipule exactement comment, techniquement parlant, les autorités peuvent s'assurer du caractère fiable et complet des données transmises. Au Royaume-Uni, la nouvelle autorité de régulation pourra exiger des rapports annuels sur les mesures mises en œuvre et le fonctionnement des algorithmes des plateformes, qui devront par ailleurs rendre leurs données accessibles aux chercheurs sous certaines conditions. Le code de conduite de la Commission Européenne enfin, incite les plateformes a collaboré avec les Etats-Membres afin de leur livrer les informations demandées par les différentes autorités de régulation.

Toutes ces mesures posent la question essentielle de l'accès aux données des plateformes. En termes d'audit des algorithmes par exemple, différentes méthodes co-existent. La rétro-ingénierie notamment, permet d'évaluer le fonctionnement d'algorithmes à travers des tests réalisés en position d'usager (Cardon, 2019) : de multiples requêtes sont réalisées afin de comparer les résultats et déterminer le fonctionnement précis de l'instrument étudié (par exemple un moteur de recherche). Une autre méthode est celle de la « saturation » (Christin, 2020), qui consiste à « épuiser » les résultats à une même requête pour en produire une typologie (par exemple pour un algorithme de recommandation). Pour autant, ces méthodes ne fonctionnent pas avec tous les types d'algorithmes, et certains audits nécessitent un accès direct aux données des plateformes. Or, dans le cadre des législations analysées dans cet article, les pouvoirs publics sont toujours dépendants des données fournies par les firmes, et ne disposent pas de moyens propres de vérifier leur fiabilité. Pour les entreprises auditées, il en résulte la possibilité d'avoir recours à une forme d' « opacité stratégique » (Ananny, Crawford, 2016), en donnant accès aux informations qu'elles souhaitent divulguer, tout en laissant dans l'ombre celles qu'elles veulent garder secrètes. A la question de savoir si une forme de régulation qui passerait par un accès direct aux données des plateformes était possible et/ou souhaitable, les professionnels avec lesquels nous avons échangé chez Google se sont montrés dubitatifs : au-delà du secret des affaires -les formules des algorithmes représentant une plus-value essentielle sur un marché concurrentiel, nos deux interlocuteurs ont mis en avant les difficultés techniques à donner accès et à comprendre, depuis l'extérieur, des architectures de systèmes d'information complexes, n'ayant pas été conçues pour répondre aux questions posées par les autorités de régulation.

« *Hard* » et « *soft regulation* » en pratique

Pour autant, publiquement, les représentants des plateformes n'expriment pas d'hostilité particulière à l'égard des tentatives de régulation des états. Au contraire, dans certains cas, ils semblent même les appeler de leurs voeux. Les déclarations de Mark Zuckerberg dans le

Washington Post en mars 2019 avaient pu surprendre[55] : le PDG de Facebook y déclarait en appeler à de nouvelles formes de régulation des contenus nuisibles sur sa plateforme, tendant la main aux pouvoirs publics pour définir de nouvelles règles. La firme avait par la suite annoncé la création d'un Conseil de surveillance[56], composé de personnalités de la société civile indépendantes de l'entreprise, bénéficiant d'un pouvoir de décision et de contrainte sur les politiques de modération de la plateforme. En février 2020, la firme est allée plus loin encore en publiant un Livre Blanc intitulé « *Online Content Regulation : Charting a Way Forward* »[57], dans lequel la vice-présidente en charge de la politique des contenus, Monika Bickert, développe un ensemble de propositions visant à incarner la « vision Facebook » de la régulation des contenus sur les plateformes. De la même façon, chez Google, Kent Walker, vice-président et directeur juridique de la firme, a pris la parole à plusieurs reprises sur le sujet : en janvier 2019[58], pour affirmer les principes que Google entendait associer à une régulation juste des contenus problématiques sur internet, puis le mois suivant, dans une note intitulée « *Smart regulation for combating illegal content* »[59], dans laquelle il prend position quant aux différentes approches de la régulation qui co-existent en Europe.

Ces déclarations ne sont pas uniquement des actes de communication visant à rassurer les états et les opinions publiques. Elles intègrent une stratégie visant à afficher une volonté de collaboration avec les autorités, tout en promouvant une certaine vision de la régulation, la moins contraignante pour leurs activités de modération, et la plus à même de protéger leur modèle économique. La stratégie peut être interprétée comme une tentative de « désamorçage » d'un éventuel conflit états/plateformes, en acceptant la régulation, à condition qu'il s'agisse *d'une certaine forme de régulation*. Dans les textes cités dans le paragraphe précédent, deux sujets sensibles crispent les débats entre états et plateformes : les amendes prévues par la NetzDG, la loi Avia et dans une moindre mesure la nouvelle législation britannique, qualifiées de « *harsh penalties on those acting in good faith* » par Kent Walker ; le délai de 24h pour retirer les contenus problématiques, « *pushing platforms to err to much on the side of removing contents* » selon Monika Bickert. La NetzDG en effet, stipule que les plateformes peuvent se voir infliger une amende de 5 millions d'euros si elles échouent à retirer des contenus signalés dans le temps imparti. De la même façon, la loi Avia, avant sa censure par le Conseil Constitutionnel, prévoyait des amendes pouvant atteindre 20 millions d'euros, ou 4% du chiffre d'affaire annuel mondial total de l'entreprise (le montant le plus élevé étant retenu). La législation britannique, en construction, prévoit également que l'autorité de contrôle de l'activité des plateformes puisse infligé des « *substantial fines* », sans pour autant que le montant des amendes ne soit fixé à ce stade.

Les plateformes ne sont pas les seuls acteurs à s'opposer à ces deux mesures. En France par exemple, la menace d'amendes associée à un délai de traitement court a généré une controverse sur les effets collatéraux néfastes que la loi Avia pourrait avoir sur la liberté d'expression (Badouard, 2020). Organisations de défense des libertés sur internet[60], Conseil

[55] Washington Post, "Mark Zuckerberg: The Internet needs new rules. Let's start in these four areas", 30 mars 2019.
[56] Facebook, « Establishing Structure and Governance for an Independent Oversight Board », *Facebook Newsroom*, 17/09/2019.
[57] Facebook, « Charting a way forward : Online Content Regulation », février 2020
[58] Google, « Oversight frameworks for content-sharing platforms », *Google Blog The Keyword*, 19/01/2019.
[59] Google, « Smart régulation for combatiing illegal content », *Google blog The Keyword*, 14/02/2019
[60] Collectif, « Proposition de loi visant à lutter contre la haine sur internet : appel collectif à préserver nos droits fondamentaux dans l'espace public numérique », 16 janvier 2020.

National du Numérique[61] et même Commission Européenne[62] se sont émus des risques de sur-censure que laissaient planer ces mesures : en cas de doute sur un contenu, les grandes firmes du web préféreraient bloquer des publications légitimes, qui n'enfreignent ni les lois ni les standards de publication, plutôt que prendre le risque de se voir infliger une amende d'un tel montant. Ces mesures ont ainsi été considérées comme portant « une atteinte à l'exercice de la liberté d'expression et de communication qui n'est pas nécessaire, adaptée et proportionnée » par le Conseil Constitutionnel, et déclarées « contraires à la Constitution »[63]. Des débats similaires sur la censure et la protection des libertés sur internet ont accompagné la publication du *Online Harms White Paper* au Royaume-Uni[64].

Ces contraintes fortes sur les dispositifs de modération des plateformes que font peser ou entendaient faire peser la NetzDG et la loi Avia font de la France et l'Allemagne les incarnations d'une approche qualifiée de « *hard regulation* » par les acteurs du secteur avec lesquels nous nous sommes entretenus. Cette approche se caractérise par la création de nouveaux délits en termes d'expression publique et vise à confier de nouvelles responsabilités aux plateformes, dont l'application est contrôlée par des agences étatiques, et qui s'articulent à un système de sanctions (notamment financières) en cas de non mise en conformité. Si cette approche est l'objet de controverses en ce qu'elle fait peser des risques sur l'exercice de la liberté d'expression sur internet, elle rappelle également que la loi n'a pas dit son dernier mot en matière de régulation des environnements numériques.

Les recherches en sciences sociales portant sur la régulation d'internet, et la régulation des contenus sur Internet, ont en effet tendance, à la suite des travaux fondateurs de Lawrence Lessig, à accorder une certaine importance au pouvoir régulateur de la technologie (De Nardis, Musiani, 2016 ; Badouard et al., 2016 ; Belli, 2016). Cette importance est souvent résumée par la célèbre formule « *Code is law* », qui exprime la manière dont les architectures et ressources techniques parviennent efficacement à encadrer les comportements sur internet. Si ce pouvoir est indéniable, en ce que la technologie constitue effectivement la source de normativité la plus contraignante dans les environnements numériques – c'est-à-dire plus contraignante que la loi, le marché et les usages (Lessig, 1999), les exemples français et allemand montrent que le pouvoir de la loi, parce qu'il implique une logique de mise en conformité du côté des plateformes, demeure important. Par crainte des pénalités infligées en cas de non-respect des nouvelles dispositions légales, les plateformes modifient leur design en intégrant de nouveaux dispositifs de modération (formulaires, boutons spécifiques), font évoluer leurs procédures d'évaluation des contenus en embauchant et en formant du personnel, rendent des comptes sur leurs pratiques à travers des rapports de transparence, ou reconnaissent de nouveaux droits à leurs usagers à travers des procédures d'appel. Ces nouvelles législations réaffirment ainsi le pouvoir régulateur des autorités publiques sur les activités des plateformes et les comportements en ligne, montrant, pour pasticher Lessig, que *la loi aussi peut faire le code.*

Pour autant, l'approche franco-allemande ne fait pas l'unanimité sur le territoire européen, et des approches qualifiées de « soft régulation » co-existent. Celles-ci ont la préférence des

[61] Conseil national du numérique, *Le CNNUM exprime ses interrogations sur la proposition de loi visant à lutter contre la haine en ligne*, Communiqué de presse du 21 mars 2019.
[62] Commission Européenne, Notification 2019/412/F, Loi visant à lutter contre les contenus haineux sur internet, 22/11/2019.
[63] Conseil Constitutionnel, art.cit.
[64] B. Haggart, N. Tusikov, "What the U.K.'s Online Harms white paper teaches us about internet regulation", *The Conversation*, 18 avril 2019.

plateformes, qui les qualifient, comme les institutions européennes d'ailleurs, de « *smart regulation* ». Ces approches reposent sur la mise en place de partenariats imposant aux entreprises du web de prendre un certain nombre d'engagements, sans toutefois prévoir de sanctions en cas de non-respect. L'approche « *hard* » contraint les plateformes à une obligation de résultats, l'approche « *soft* » leur impose « simplement » une obligation de moyens. Les mesures prises par la Commission Européenne depuis 2016 illustrent bien cette seconde approche. Rappelons ici que pour la Commission, le principal enjeu en matière de régulation des contenus réside dans l'harmonisation des législations nationales. Or, en matière de désinformation et de discours de haine, l'exposition des populations européennes à ces problèmes, comme les réponses apportées par les états, diffèrent grandement. Au-delà de son rapport aux plateformes, la Commission a donc intérêt à adopter une certaine souplesse dans l'application de ses stratégies de régulation.

En mai 2016, la Commission a conclu un partenariat avec Facebook, Microsoft, Twitter et YouTube, baptisé « Code de conduite pour la lutte contre les discours haineux illégaux en ligne »[65]. Ce code de conduite va inspirer une communication de la Commission sur la lutte contre les discours haineux en ligne, publiée en septembre 2017[66]. Les entreprises impliquées dans ce partenariat, parmi lesquelles on retrouve les principales plateformes de réseaux sociaux, ont pris un certain nombre d'engagements. On retrouve par exemple la clarification de leurs standards de publication auprès des usagers, la mise en place de procédures de signalement, l'embauche d'équipes de modérateurs, le retrait des contenus notifiés en moins de 24 heures, la collaboration avec des organisations de la société civile ou encore le respect des différentes législations nationales sur le sujet. Pour faire respecter ces règles, la Commission a adopté une stratégie de *testing*, qui vise à évaluer les procédures mises en place par des acteurs de la société civile.

La méthode est la suivante : des associations de lutte contre le racisme, partenaires de la Commission, signalent aux plateformes qui ont signé le code de bonne conduite un ensemble de contenus manifestement racistes, antisémites, misogynes ou homophobes, en se faisant passer pour des utilisateurs lambda, ou en assumant leur rôle d'association partenaire (appelées « *trusted flaggers* »[67]). Elles recensent ensuite les réponses apportées par les plateformes et mesurent les délais de traitement de leurs signalements. La première évaluation, réalisée à l'automne 2016[68], portait sur 600 signalements réalisés par 12 organisations basées dans 9 pays différents. La dernière évaluation, réalisée en février 2019[69], a porté sur plus de 4000 signalements et a impliqué 39 organisations venant de 26 pays différents. Le principal résultat de cette évaluation est que si 72 % des contenus notifiés ont été retirés (dont 89 % en moins de 24 heures), il subsiste une grande disparité entre les plateformes. Facebook et YouTube retirent par exemple plus de 80 % des contenus notifiés, alors que Twitter, réputé plus réticent en matière de modération, n'en retire que 43 %. Les internautes qui signalent des contenus reçoivent des retours de la plateforme dans 93 % des cas sur Facebook, contre seulement 25 % des cas sur YouTube. Dans tous les cas, les

[65] Commission Européenne, *Code of Conduit on countering illegal hate speech online*, 30 juin 2016.
[66] Commission Européenne, *Communication on Tackling Illegal Content Online - Towards an enhanced responsability of online plateforms*, 28/09/2017.
[67] Les *trusted flaggers*, ou rapporteurs de confiance, sont des associations spécialisées dans la défense des droits des personnes ou la lutte contre les discriminations, partenaires des plateformes, qui disposent de canaux spécifiques qui leur garantissent de voir traiter leurs signalements en priorité.
[68] Les différents rapports d'évaluation sont disponibles à l'adresse suivante : https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en
[69] Commission européenne, *4th monitoring round of the Code of Conduct*, février 2019.

rapporteurs de confiance voient leurs signalements bénéficier d'une meilleure prise en compte et disposent de davantage de retour de la part des plateformes que les internautes ordinaires.

Si la Commission européenne se satisfait de ces résultats, c'est d'abord parce que ceux-ci se sont nettement améliorés depuis la première évaluation. En 2016, seulement 28 % des signalements avaient été pris en compte. Le pourcentage de retrait est passé de 28 à 82 % sur Facebook, de 19 à 43 % sur Twitter et de 48 à 85 % sur YouTube. Par ailleurs, un des principaux résultats des premières évaluations était également la mise en évidence d'une grande disparité suivant les pays européens, qui tend à se résorber dans la dernière évaluation, puisque tous les pays de l'Union européenne, à l'exception de la Finlande et du Portugal, passaient la barre des 50 % de retrait. La stratégie d'homogénéisation des pratiques de modération à l'échelle européenne semble ainsi porter ses fruits, et préfigure les discussions autour du *Digital Service Act*, qui doit mettre à jour la directive e-commerce de 2000 avant la fin de l'année 2020 en matière, notamment, de responsabilité des plateformes et de droit des internautes.

Le Royaume-Uni, de son côté, développe une approche à la croisée de la « soft » et de la « hard » régulation. La principale mesure du *Online Harms White Paper* publié en avril 2019[70], réside dans la création d'une nouvelle autorité publique indépendante, chargée de fixer un code de conduite concernant les pratiques de modération sur les réseaux sociaux, et ayant pour mission de vérifier que les plateformes s'y conforment. La logique déployée dans le document est celle d'un « *duty of care* », soit une obligation de moyens : l'autorité publique demande aux plateformes de mettre en œuvre des mesures et contrôle que celles-ci soient bien mises en œuvre, sans imposer aux acteurs du numérique des objectifs chiffrés. Pour autant, l'autorité en question, financée par un impôt sur les secteurs de télécommunications et du numérique, doit également bénéficier d'un pouvoir de sanction sur les plateformes, notamment en leur infligeant des amendes, si les engagements ne sont pas tenus.

Ce type d'approche mixte, associant partenariat, obligation de moyens, et pouvoirs de contrainte en cas de non-respect des engagements, avait également été exploré en France à travers la mission Loutrel. Annoncée par le président Emmanuel Macron lors de la tenue de l'*Internet Governance Forum* à Paris en 2018[71], la mission consistait à mandater une équipe gouvernementale, accueillie par Facebook lors d'une mission d'observation des dispositifs de modération, afin de formuler conjointement un ensemble de propositions pour une régulation plus « efficace » des contenus. Le rapport remis par la mission[72] défendait une approche d' « *accountability by design* », se traduisant par une obligation de moyens, tout en prévoyant que les autorités administratives en charge du contrôle des activités des plateformes bénéficient d'un pouvoir d'audit, en bénéficiant d'un accès direct à leurs données. Finalement, les arbitrages gouvernementaux ont donné raison à l'équipe réunie autour de la députée Laëtitia Avia, mettant fin prématurément à la mission Loutrel[73]. La censure de la loi Avia par le Conseil Constitutionnel en juin 2020 laisse présager un retour de l'approche de la « *soft regulation* » en France : en juillet 2020, la publication d'une tribune commune par le

[70] Gouvernement du Royaume-Uni, *Online Harm White Paper*, avril 2019.
[71] Elysée, *Discours du Président de la République, Emmanuel Macron lors du forum sur la gouvernance de l'internet à l'UNESCO*, 12/11/2018
[72] Secrétariat d'Etat en charge du numérique, *Créer un cadre français de responsabilisation des réseaux sociaux : agir en France avec une ambition européenne*, Rapport de la mission « Régulation des réseaux sociaux - Expérimentation Facebook », mai 2019
[73] Elisa Braün, « How Macron tried to fix Facebook – and failed », *Politico*, 21/10/2019.

responsable du CSA et de son homologue allemand quant au *Digital Service Act* semblait aller dans ce sens[74].

Typologie des formes de régulation des contenus

	Cas étudiés	Mesures	Sanctions
« Hard regulation »	Loi Avia NetzDG	Obligation de résultats (définition d'objectifs et de délais strictes)	Amendes en cas d'échec à atteindre les résultats ou à respecter les délais
« Soft regulation »	Partenariat européen	Obligation de moyens (définition de mesures à mettre en œuvre)	/
Approche intermédiaire	Online Harms White Paper Loi sur les manipulations de l'information	Obligations de moyens Audit des plateformes et/ou contraintes légales de mise en oeuvre	Amendes (en cas de non mise en œuvre de moyens)

Conclusion

L'analyse de l'évolution des politiques de modération de Facebook et YouTube en Europe tend à contredire la supposée impuissance des pouvoirs publics à réguler les environnements numériques. Dans l'étude présentée dans cet article, on constate à l'inverse que les nouvelles normes imposées par les états poussent les plateformes à adopter une logique de mise en conformité, qui se traduit notamment par une évolution des procédures de signalement et d'évaluation des contenus signalés. Les approches relevant de la « *hard regulation* », notamment en Allemagne, semblent particulièrement efficaces à infléchir les stratégies de régulation des plateformes. Pour autant, même celles relevant de la « soft regulation », comme le partenariat mis en place par l'Union Européenne, produisent des effets sur les activités de modération des firmes du web, qui intensifient les traitements des signalements. S'il est difficile de statuer sur *les raisons* qui poussent des acteurs comme Facebook et YouTube à réformer leurs dispositifs de modération, notamment lorsque ceux-ci ne sont pas exposés à des sanctions (objectifs économiques, pression sociale, respect des engagements, stratégie de communication, etc.), il n'en demeure pas moins que les politiques des plateformes en la matière ont connu de profondes révisions ces dernières années.
Cette réforme s'articule autour d'un double mouvement : une politique d'invisibilisation des contenus problématiques d'une part, et une logique d'automatisation de la modération de l'autre. L'invisibilisation, qui consiste à jouer sur les infrastructures de l'attention (algorithmes de tri et de recommandation notamment) pour accorder des niveaux de visibilité différenciés aux publications, présente à la fois une certaine efficacité pour limiter la viralité des fausses informations et des contenus haineux, tout en incarnant un risque de censure abusive lié à l'opacité des mécanismes en question. Par ailleurs, cette logique consacre

[74] Conseil Supérieur de l'Audiovisuel, *Tribune commune de Roch-Olivier Maistre et Tobias Schmid en faveur d'une nouvelle régulation européenne*, 09/08/2020.

paradoxalement le rôle proprement éditorial des plateformes, qui filtrent et hiérarchisent les publications en fonction de la qualité de leurs contenus, ce qu'elles s'étaient toujours (officiellement) refusées à faire depuis leur création. L'automatisation, qui vise à détecter mécaniquement des contenus problématiques pour faire face à l'important volume de contenus publiés, pose de son côté un double problème d'efficacité et d'erreurs d'évaluation. Face à ces écueils, certaines plateformes, comme Facebook et YouTube, développent des procédures d'appel, qui permettent aux internautes dont les *posts* ont été supprimés d'exiger un réexamen de leurs publications.

La réponse des pouvoirs publics en Europe, au double mouvement d'automatisation et d'invisibilisation, réside principalement dans une injonction à la transparence, qui consiste à imposer aux plateformes la publication de rapports réguliers et la mise en place de dispositifs d'information à destination des internautes. Cette injonction à la transparence, si elle permet de mieux faire connaître les activités des firmes et de les soumettre à une critique publique, pose également la question majeure de l'accès aux données, les pouvoirs publics ne bénéficiant pas, *in fine*, de moyens de certification des informations qui leur sont fournies. En conclusion donc, si le cas d'étude des politiques de modération des plateformes invite à ré-évaluer le pouvoir normatif des pouvoirs publics dans les environnements numériques, la volonté politique des acteurs étatiques se heurte à ces enjeux d'accès, de vérification et de certification, dont les modalités hypothétiques de mise en œuvre (audit, contrôle judiciaire, accès direct aux données), demeurent encore floues.

Pour terminer, notons que l'étude présentée dans cet article comporte trois limites qu'il paraît important de souligner. La première est qu'elle se concentre sur le cas des « fausses informations » et des discours de haine, alors que la question de la régulation des contenus sur internet se joue également dans le cadre d'autres dossiers d'importance, notamment la lutte contre la propagande djihadiste et la protection du droit d'auteur (Tréguer, 2019). La seconde est qu'elle passe sous silence tout ce qui a trait au travail des modérateurs, au marché de la modération, et aux controverses dont ils sont l'objet. Les projets de régulation étudiés dans cet article ne prennent pas en charge ces thématiques, qui nous ont dès lors paru hors sujet pour cette étude. Je renvoie les lecteurs et lectrices intéressé.e.s aux travaux de Sarah Roberts et Tarleton Gillepsie sur le sujet. La troisième limite est de ne se concentrer, volontairement, que sur les relations entre états et plateformes, alors même que les pratiques de régulation des contenus englobent une grande diversité d'acteurs, au rang desquels on trouve des organisations de la société civile et des médias, des régies publicitaires et des annonceurs, des experts et des internautes ordinaires (Badouard, 2020). Ces acteurs incarnent, à différents niveaux, la pluralité des pouvoirs normatifs à l'œuvre sur le web. Parce que les nouveaux régimes de gouvernement de la parole sur les réseaux sociaux bousculent l'exercice de la liberté d'expression sur internet, une régulation réellement démocratique des contenus passe par l'équilibre entre ces différents pouvoirs.

Références bibliographiques

ANANNY M., CRAWFORD K. (2016), "Seeing without knowing : Limitations of the transparency ideal and its application to algorithmic accountability", *New Media & Society*, 20 (3), p.973-989.

BADOUARD R. (2020), *Les nouvelles lois du web. Modération et Censure*, Coll. La République des Idées, Le Seuil.

BADOUARD R., MABI C., SIRE G. (2016), "Beyond "Points of Control": Logics of Digital Governmentality", *Internet Policy Review*, vol.5, n°3.

BARZILAI-NAHON N. (2008), "Toward a theory of network gatekeeping: A framework for exploring information control", *Journal of the American Society for Information Science and Technology*, 59/9, 2008, p. 1493-1512.

BELLI L. (2016), *De la gouvernance à la régulation de l'internet*, Berger Levrault.

CARDON D. (2019), « Auditer les algorithmes », in *Cultures numériques*, Presses de Sciences Po, p. 399-409.

CHEVRET-CASTELLANI C., LABELLE S. (2019), « Transparence et loyauté, deux motifs de régulation des algorithmes », *Terminal*, n°124.

CHRISTIN A. (2020), "The ethnographer and the algorithm: beyond the black box", *Theory and Society*, Online First, p.1-22.

DE NARDIS L., MUSIANI F. (2016), « Governance by Infrastructure », in MUSIANI et al. (dir), *The Turn To Infrastructure in Internet Governance*, Palgrave MacMillan, p.3-21.

FADDOUL M., CHASLOT G., FARID H. (2020), "A longitudinal analysis of YouTube's promotion of conspiracy videos", pre-print disponible à l'adresse (page visitée le 4/11/2020) :

https://arxiv.org/abs/2003.03318

GILLEPSIE T. (2018), *Custodians of the Internet. Plateforms, Content Moderation, and the Hidden Decisions That Shape Social Media*, Yale University Press.

GRIMMELMANN J. (2014), "Speech Engines", U. of Maryland Legal Studies Research Paper n°2014-11.

LESSIG L. (1999), *Code and Other Laws of Cyberspace*, New-York, Basic Books.

MAGGETTI M. (2015), "Hard and Soft Governance", in Lynggaard K., Manners I., Löfgren K. (dir.), Research Methods in European Union Studies, Palgrave MacMillan, p.252-265.

MATTELART T. (2020), « Comprendre la stratégie de Facebook à l'égard des médias d'information », *Sur le journalisme*, vol.9 n°1, p.24-43.

MYERS WEST S. (2017), "Raging Against the Machine: Network Gatekeeping and Collective Action on Social Media Plaftorms", *Media and Communication*, vol.5, n°3, p.28-36.

ROBERTS S.T. (2019), *Behind the Screen. Content Moderation in the Shadows of Social Media*, Yale University Press.

SCHAFER V. (2018), *En construction. La fabrique française d'internet et du Web dans les années 1990*, Paris, INA Editions.

SIRE G. (2015), « Cinq questions auxquelles Google n'aura jamais fini de répondre », *Hermès*, n°73, p. 201-208.

TREGUER F. (2019), *L'utopie déchue. Une contre-histoire d'Internet. XV-XXIème siècle*, Paris, Fayard.

Résumé

La modération sur les réseaux sociaux a été l'objet de nombreuses controverses ces dernières années. Face à la pression des états européens pour inciter les grandes plateformes à réguler davantage les contenus qu'elles hébergent, celles-ci ont entrepris une réforme de leurs

politiques de modération. A partir des exemples de YouTube et Facebook, trois aspects de cette réforme sont étudiés dans cet article : la révision des procédures de signalement, l'automatisation de la modération et la régulation par les infrastructures, qui consiste à jouer sur la visibilité des publications pour limiter leur viralité. A travers l'analyse des interactions entre états et plateformes sur les dossiers de la régulation des fausses informations et des discours de haine, cet article entend contribuer à la caractérisation des nouveaux régimes de gouvernement de la parole publique sur les réseaux sociaux.

Mots clés

Modération ; régulation ; plateformes ; réseaux sociaux ; fausses informations ; discours de haine.