UQÀM

Centre d'études sur l'intégration et la mondialisation

CEIM

La lutte contre les contenus haineux sur les plateformes de médias sociaux : une analyse comparative d'approches de régulation

Note de recherche, Dorian Mouketou, 9 juillet 2021

Résumé

Notre recherche traite de la réglementation des plateformes de médias sociaux, notamment dans la lutte contre les discours haineux. Elle s'appuie sur l'analyse comparative de deux approches de régulation adoptées dans deux juridictions distinctes : la *soft regulation* (Union européenne) et la *hard regulation* (Allemagne). Malgré les variations de contrainte des deux approches, nous constatons qu'elles ne se traduisent pas en suppressions excessives de contenus répréhensibles par les médias sociaux (par souci de conformité aux règles), et ce, malgré une augmentation du volume de signalements ou de plaintes reçues. Les plateformes se sont généralement conformées à leurs obligations, alors que les mesures de régulation ont favorisé davantage de transparence sur leurs pratiques de modération.

Mots-clés : régulation, plateformes en ligne, médias sociaux, préjudices en ligne, discours haineux.

Abstract

Our research discusses the regulation of social media platforms, particularly in the fight against hate speech. It is based on a comparative analysis of two regulatory approaches adopted in two distinct jurisdictions: soft regulation (European Union) and hard regulation (Germany). Despite the variations in the constraints of the two approaches, we find that they do not result into excessive deletions of objectionable content by social media platforms (for the sake of compliance with the rules), despite an increase in the volume of reports or complaints received. The platforms have generally complied with their obligations, while the regulatory measures have fostered more transparency on their moderation.

Keywords: regulation, platforms, social media, online harms, hate speech.

Contenu

Introduction[1]

La liberté d'expression vaut non seulement pour les informations ou les idées accueillies avec faveur, ou considérées comme inoffensives ou indifférentes, mais aussi pour celles qui heurtent, choquent ou inquiètent l'État ou une fraction quelconque de la population. Ainsi le veulent le pluralisme, la tolérance et l'esprit d'ouverture sans lesquels il n'y a pas de société démocratique.

Cour européenne des droits de l'homme (CEDH), 1976.

La montée des crimes haineux semble coïncider avec l'exacerbation des propos haineux sur les médias sociaux (Armstrong 2019). En effet, un rapport publié par Statistique Canada en 2019 et exposant les données sur les crimes haineux montrait que les crimes motivés par la haine avaient augmenté de 47 % entre 2016 et 2017[2]. Cette hausse serait attribuable à l'augmentation de crimes haineux à l'égard d'une religion, d'une race ou d'une origine ethnique. Plus précisément, entre 2016 et 2017, le nombre de crimes motivés par la haine à l'égard d'une race ou d'une origine ethnique s'est accru de 32 %, alors que celui des crimes motivés par la haine à l'égard d'une religion s'est accru de 87 % (Armstrong 2019). En 2017, une étude commandée par CBC avait également relevé une explosion de propos intolérants et racistes sur les réseaux sociaux au Canada depuis la course à la Maison-Blanche de Donald Trump. La proportion de propos racistes tenus par les Canadiens via les médias sociaux, les blogues et les sections de commentaires en ligne avait grimpé de 600 % entre octobre 2015 et novembre 2016 (« Canadians appear to be more hateful online. Here's what you can do about it » 2017: par. 3; Robillard 2017; Housefather 2019)[3].

Dans la foulée de ces révélations, le Comité permanent de la justice et des droits de la personne de la Chambre des Communes du Canada avait décidé de se pencher sur l'étude de la haine en ligne au Canada, en mars 2019 (Housefather 2019). À l'issue de leurs travaux, les députés avaient formulé certaines recommandations pour combattre la haine en ligne. Plus récemment, dans un rapport intitulé « L'avenir des télécommunications : le temps d'agir », le Groupe d'examen du cadre législatif en matière de radiodiffusion et de

[1] Publication tirée du rapport de stage de fin de maitrise en administration publique de l'étudiant. L'étudiant tient à remercier le professeur Stéphane Paquin d'avoir accepté de diriger son rapport de stage et de lui avoir prodigué des conseils qui ont orienté de façon positive l'angle d'analyse dans le rapport.

[2] En 2017, le nombre de crimes motivés par la haine était évalué à 2 073, contre 1 409 en 2016 et 1 364 en 2015. Il convient de noter qu'en 2016, ce nombre représentait moins de 0,1 % des 1 895 546 crimes. En 2017, ce nombre représentait 0,1% des 1 940 846 crimes. Ainsi, malgré une hausse du nombre de crimes motivés par la haine, celle-ci occupe une proportion autour de 0.1 % du nombre total de crimes (Voir Allen 2019; Statistique Canada 2017 et 2018).

[3] La vérification de l'échantillon de cette étude demeure difficile puisqu'elle cette dernière n'est pas publique ou disponible. Il est toutefois indiqué que la firme de recherche Cision a utilisé un « échantillon représentatif de 1 % de mentions sur Twitter, de forums en ligne, de blogues et de sites web de médias [...] pour retracer 40 termes jugés racistes ou dégradants envers la religion ou l'origine ethnique d'une personne » (Robillard 2017 : par.10).

télécommunications, présidé par la juge Janet Yale, s'était également penché sur la question des contenus préjudiciables en ligne. Le Groupe avait recommandé au gouvernement du Canada d'adopter une « mesure législative en matière de responsabilité aux fournisseurs de services numériques pour les contenus et comportements préjudiciables utilisant les technologies numériques, au-delà de toute responsabilité imposée par les lois des communications » (Yale 2020 : 216). En 2020, le gouvernement du Canada rendait publique une note pour la période des questions visant à annoncer une éventuelle « réglementation des plateformes de médias sociaux » (Gouvernement du Canada 2020b), pilotée par le ministre du Patrimoine canadien. Cette publication confirmait l'engagement du gouvernement communiqué dans la lettre de mandat du ministre, mais aussi stipulé dans le discours du Trône prononcé le 23 septembre 2020, afin de prendre des mesures pour lutter contre la haine en ligne (Premier ministre du Canada 2021; Payette 2020). Depuis le début de son mandat, le ministre du Patrimoine, Steven Guilbeault, indique également dans les médias qu'un projet de loi fixant un nouveau cadre de régulation des plateformes sera bientôt déposé à la Chambre des communes (Castonguay 2021; Noël 2021; Silver 2021).

Or, la régulation des propos haineux sur les médias sociaux pose le problème du « juste équilibre entre le maintien d'un espace libre et ouvert pour l'échange d'idées et d'information, le respect et la protection des droits et libertés individuels et collectifs » (Yale 2020 : 15). En guise d'illustration, les réactions qui ont fusé à travers le monde entier après la suspension des comptes de Donald Trump témoignent de cette complexité, ravivant les débats sur la liberté d'expression et la régulation des médias sociaux. En France, par exemple, une grande partie de la classe politique, tous partis confondus, a condamné la démarche de Facebook, de Twitter et de YouTube, y voyant une tentative des plateformes de s'arroger le pouvoir de décider du contenu en ligne (Piquard 2021). Alors que la question divise (Krishnamurthy 2021), il semblerait toutefois que l'opinion publique soit favorable à une telle législation. En effet, Radio-Canada révélait qu'un sondage mené par la firme Abacus Data pour le compte de la Fondation canadienne des relations raciales (FCRR) – une institution ayant pour mission la lutte contre le racisme au Canada – montrait que 60 % des Canadiens étaient en faveur d'une réglementation pour lutter contre « la diffusion de discours et de comportements haineux et racistes en ligne (Martinez 2021 : par.1).

Toutefois, la régulation des géants numériques est une entreprise complexe. Néanmoins, il semble y avoir un *momentum* à l'échelle internationale pour davantage de réglementation, y compris aux États-Unis. Ces dernières années, des tentatives se sont multipliées entre autres pour répondre à la domination et au déséquilibre de pouvoirs entre les géants numériques et les plus petites entreprises, dont les médias traditionnels, pour lutter contre la désinformation et les contenus illégaux en ligne, ou encore pour assurer que les géants numériques paient de l'impôt dans les pays où ils opèrent. En ce qui a trait notamment à la lutte contre les discours haineux en ligne, plusieurs juridictions ont déjà mis en place ou tentent de mettre en place des mesures, soit pour impliquer davantage les plateformes de médias sociaux dans la lutte et la prévention contre les discours de haine, soit pour les contraindre à se conformer aux lois nationales.

Citons en exemple l'Allemagne, qui a été un des premiers pays à avoir adopté une loi, en 2017, pour renforcer l'application des dispositions relatives aux contenus illégaux (*Code*

criminel) sur les médias sociaux. La France a tenté la même expérience en 2020, mais sa loi visant à lutter contre la haine en ligne a été invalidée par le Conseil constitutionnel. Au niveau supranational, la Commission européenne (CE) a mis en place en 2016 un code de conduite en partenariat avec des plateformes majeures pour lutter contre les discours haineux illégaux en ligne. Cette année, l'Inde a adopté des règles hautement controversées visant à réguler les contenus sur les réseaux sociaux, mais aussi sur les services de messagerie privée. D'autres pays sont également en train de proposer des mesures en ce sens. Pensons à l'éventuel projet de loi sur la sécurité en ligne et la réglementation des médias en Irlande, au projet de loi sur la sécurité en ligne en Australie, ou encore au Livre blanc sur les préjudices en ligne du Royaume-Uni de 2019, qui constitue un nouveau cadre réglementaire en préambule d'un éventuel projet de loi sur la sécurité en ligne.

Bref, le contexte actuel propulse l'enjeu de la lutte contre les contenus préjudiciables en ligne, y compris les discours de haine, dans l'agenda du gouvernement du Canada. En politique publique, l'analyse des solutions retenues ailleurs peut s'avérer une étape importante pour collecter des informations sur les meilleures pratiques (Knoepfel Larrue, Varone et Savard 2015). Cette approche d'analyse guidera notre recherche, qui se penchera sur un corpus de connaissances relatif à la régulation des contenus sur les plateformes en ligne, mais aussi sur une analyse comparative des initiatives ayant déjà été mises en place dans d'autres juridictions.

Les plateforme numériques : un phénomène complexe à élucider

Le concept de « plateforme » renferme une signification ambiguë, car sa définition relève différentes variations selon les champs de connaissances. Les informaticiens, les économistes, les spécialistes des médias numériques, et même les avocats, ont tous différentes approches pour aborder ce terme (Gillespie 2010; Flew, Martin et Suzor 2019; Gorwa 2019b; Gasser et Schulz 2015). Néanmoins, pour formuler une définition synthétique, il est possible de définir les plateformes comme des structures et infrastructures sociotechniques permettant et facilitant l'interaction et la communication entre les acteurs économiques (utilisateurs), notamment par la collecte, la circulation et la diffusion des données des utilisateurs (Helberger, Pierson et Poell 2018; Flew, Martin et Suzor 2019; Gillespie 2018a). Le développement des plateformes fait ainsi référence au phénomène d'intermédiation en ligne via un espace ou un service (plateforme numérique) qui permet aux utilisateurs d'interagir (Cordier 2019).

Par ailleurs, les termes « intermédiaire » et « plateforme » tendent à semer la confusion, notamment pour leur utilisation tantôt simultanée, tantôt alternée, mais aussi, dans certaines circonstances, leur utilisation croisée (par exemple « plateformes d'intermédiation »). L'Organisation de coopération et de développement économiques (OCDE) (2010 : 9) a tenté de proposer une définition de ce que sont les intermédiaires en ligne : « Internet intermediaries bring together or facilitate transactions between third parties on the Internet. They give access to, host, transmit and index content, products and services originated by third parties on the Internet or provide Internet-based ser-

vices to third parties ». Selon Thelle, Sunesen, Basalisco, la Cour Sonne et Fredslund (2015), cette définition n'est pas claire, compte tenu du contexte évolutif et changeant des intermédiaires en ligne. Néanmoins, elle a le mérite de souligner deux importantes fonctions ou caractéristiques intrinsèques aux intermédiaires en ligne. *Primo*, les intermédiaires en ligne facilitent les interactions entre d'autres personnes (tiers). *Secundo*, le contenu étant produit par ces tiers, les plateformes sont donc des sites et des services en ligne qui ne produisent ou ne commandent pas l'essentiel du contenu qu'ils hébergent ou diffusent (Gillespie 2017).

En effet, comme le stipule l'OCDE (2010), le terme « intermédiaire » a un sens implicite puisqu'il est situé « between or among two or more parties, and although they help in the transmission/dissemination process, intermediaries do not initiate decisions to disseminate the content, products or services that transverse their networks or servers ». Les intermédiaires peuvent être classés sous différentes catégories, à savoir : les fournisseurs d'accès et de services Internet (FAI), les fournisseurs de traitement de données et d'hébergement Web, les moteurs et les portails de recherche Internet, les plateformes de commerce électronique (E-Commerce), les systèmes de paiement sur Internet et les plateformes de réseautage dit « participatif » (OCDE 2010). Les plateformes de médias sociaux, qui nous intéressent particulièrement, peuvent être incluses cette dernière catégorie d'intermédiaires.

Les plateformes de médias sociaux

Les plateformes de médias sociaux ont pour principale fonction de faciliter la communication sociale et le partage d'informations entre les utilisateurs (OCDE 2010; Bakalis et Hornle 2021). Selon Abar et Wildman (2015), les médias sociaux partagent quatre caractéristiques. D'abord, ce sont des applications Internet interactives. En effet, l'avènement des applications Web 2.0 a permis le passage d'un média de consommation, permettant de lire ce que les autres écrivaient ou de consommer des audios et des vidéo-clips, à un média d'interaction entre les utilisateurs et favorisant une logique de participation, où les utilisateurs peuvent créer et partager du contenu avec les autres. Aussi, le contenu généré par les utilisateurs est la pierre angulaire des médias sociaux : « Web 2.0 is the ideology and user-generated content is the fuel » (Abar et Wildman 2015 : 7). Sans le contenu des utilisateurs, les médias sociaux seraient des villes fantômes. Une troisième caractéristique qui constitue la colonne vertébrale des médias sociaux est l'existence du profil d'utilisateur, qui permet aux plateformes non seulement de stocker de l'information, mais aussi de connecter les utilisateurs entre eux. Enfin, les médias sociaux facilitent le développement de réseaux sociaux (connexions) en ligne (Abar et Wildman 2015).

Comme l'indique Gozlan (2013 : 121), « [c]e qui est en jeu dans les médias sociaux est la notion de partage; il s'agit de se retrouver autour d'un intérêt commun ou d'un lien dit d'amitié qu'il soit véritable, professionnel, superficiel ». Ainsi, les médias sociaux comme Facebook, par exemple, se distinguent par leur logique de « communautés virtuelles » (Gozlan 2013). Les plateformes de médias sociaux comprennent une variété de catégories, à savoir les sites de réseautage social – ou réseaux sociaux (Facebook, LinkedIn, etc.), les sites de partage de photos et d'images (Instagram, Flickr, Snapchat, etc.), les sites de blogues et de microblogage (Twitter, WordPress, etc.), les sites de partage de vidéos (YouTube, Vimeo, DailyMotion), les sites de discussion (Reddit), les sites de diffusion en

direct (Facebook Live, périscope), etc. (OCDE 2010; Flew, Martin et Suzor 2019; Cordier 2019; Gozlan 2013).

Le mythe de la neutralité des plateformes de médias sociaux

Les plateformes de médias sociaux se définissent comme des intermédiaires. Comme nous l'avons déjà indiqué, la définition des intermédiaires fournie par l'OCDE (2010) a deux implications : d'une part, les intermédiaires facilitent les interactions entre les utilisateurs; d'autre part, le contenu produit sur la plateforme est généré par ces mêmes utilisateurs. Cette seconde caractéristique implique – en théorie – que les intermédiaires n'interviennent pas dans la publication de contenus, agissant simplement comme des fournisseurs de services d'intermédiation; ce qui ferait d'eux des acteurs neutres. Or, nous le verrons, ce positionnement ne fait guère consensus.

Bien que considérés neutres – du moins sur le plan technique – dans le sens qu'ils sont des plateformes d'intermédiation de contenus publiés et partagés par les utilisateurs, il n'en demeure pas moins que les médias sociaux prennent des décisions relatives aux contenus autorisés sur leurs plateformes et établissent des critères selon lesquels certains contenus doivent être supprimés (DeNardis et Hackl 2015). En ce sens, ils exercent un pouvoir direct sur les droits des utilisateurs en ligne. Ces plateformes jouent un double rôle « as both a private company and a public space playing a pivotal role as access points to information » (Jørgensen et Zuleta 2020 : 2), étant une sphère privée et juridiquement un service commercial, libre de définir ce qui est autorisé et ce qui ne l'est pas. Selon Gillespie (2018b), les plateformes de médias sociaux seraient dans une catégorie hybride entre les canaux d'information interpersonnels classiques, comme les compagnies de téléphone, et les fournisseurs de contenu médiatique, tels que la radio, le cinéma, les magazines, les journaux, la télévision ou encore les jeux vidéo.

> But as a part of their service, these platforms not only host that content, they organize it, make it searchable, and in some cases even algorithmically select some subset of it to deliver as front-page offerings, news feeds, subscribed channels, or personalized recommendations. In a way, those choices are the central commodity platforms sell, meant to draw users in and keep them on the platform, in exchange for advertising and personal data. Users entrust platforms with their interpersonal "tele-" communication, but those contributions then serve as the raw material for the platforms to produce an emotionally engaging flow, more like a "broadcast" (Gillespie 2018b: 209).

Les plateformes ne sont pas que des hébergeurs de contenus, mais constituent également des acteurs essentiels, notamment dans leurs efforts d'amélioration de l'engagement des utilisateurs et la diffusion virale du contenu. Cette réalité les placerait dans une position ambiguë : d'une part, il y a une incitation économique à favoriser la non-discrimination et la libre circulation des contenus en ligne; alors que, d'autre part, ces plateformes n'ont pas intérêt à s'aliéner leur plus large base d'utilisateurs, ce qui les pousse à favoriser des pratiques de signalement et de suppression de contenus problématiques ou préjudi-

ciables (Helberger, Pierson et Poell 2018). En ce sens, pour Gillespie (2017, 2018a, 2018b), l'impartialité des plateformes est un mythe :

> [T]hey are designed to invite and shape participation, toward particular ends. This includes what kind of participation they invite and encourage; what gets displayed first or most prominently; how the platforms design navigation from content to user to exchange; the pressures exerted by pricing and revenue models; and how they organize information through algorithmic sorting, privileging some content over others, in opaque ways (Gillespie 2017: 5).

Le rôle que jouent les plateformes sur le contenu est également un enjeu en Europe où la directive sur le commerce électronique stipule que ces acteurs sont exemptés de responsabilité pour les contenus qu'ils stockent ou hébergent, et ce, s'ils agissent de manière strictement passive (E-Commerce Directive 2020). Or, il n'en demeure pas moins que la question de savoir si les plateformes agissent de manière active ou passive suscite des débats en Europe, la ligne de démarcation étant mince (Martens 2016).

Selon Gillespie (2017), les plateformes sont confrontées à trois défis distincts qui révèlent les limites du régime de responsabilité (*liability regime*) : *primo*, la plupart des lois établissant les sphères de sécurité n'ont pas été pensées en tenant compte des plateformes de médias sociaux; *secundo*, alors que les normes de responsabilité des intermédiaires sont généralement spécifiques à un pays, les plateformes sont généralement établies dans plusieurs pays, où se trouvent leurs utilisateurs, mais aussi des juridictions aux logiques différentes; *tertio*, en raison des contenus particulièrement odieux circulant sur leurs plateformes, les présomptions sur le niveau de responsabilité des plateformes sont de plus en plus remises en question par les utilisateurs et les gouvernements. Cela peut sans doute expliquer la pression grandissante publique et la croissance des législations actuelles visant à réguler davantage les contenus en ligne.

Les plateformes de médias sociaux et les discours haineux

Il n'existe pas de définition universelle de ce qu'est un discours haineux (Banks 2010; Alkiviadou 2019; Quintel et Ullrich 2020). Pour Quintel et Ullrich (2020), cela s'explique, d'une part, par la variabilité de l'interprétation de la liberté d'expression entre les pays ou les régions du monde, mais aussi, d'autre part, par les différenciations interdépendantes dans la conceptualisation du préjudice. Le manque de consensus universel sur ce qui est préjudiciable ou inapproprié complique ainsi la lutte contre la haine en ligne dans ce domaine (Alkiviadou 2019).

Malgré l'absence d'une définition universellement adoptée, des institutions, des chercheurs, et même des opérateurs de plateformes ont tenté de proposer une définition de ce que constitue un discours haineux, certaines de ses caractéristiques étant plutôt partagées. Chetty et Alathur (2018 : 110) soulignent, par exemple, que le discours de haine est plus destructeur et dangereux lorsqu'il cible un symbole traditionnel, un événement ou une activité : « The messages exchanged on individuals related to nation, race, ethnicity,

religion, sexual orientation, occupation, gender or disability have a more impact than the individuals personal information ». Almagor (2011 : 1) a défini le discours de haine comme un « bias-motivated, hostile, malicious speech aimed at a person or a group of people because of some of their actual or perceived innate characteristics ».

Une des définitions les plus populaires est celle fournie par la Cour européenne des droits de l'Homme, qualifiant le discours haineux comme étant :

> all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin (Council of Europe 1997 : 107).

La notion de l'identité ou des caractéristiques de la personne (ou groupe de personnes) ciblée par le discours haineux tend à ressortir de toutes ces définitions. Selon Chetty et Alathur (2018), le discours haineux peut ainsi cibler une personne ou un groupe de personnes sur la base du sexe, de la religion, de la race et du handicap :

- Discours haineux sexiste (*Gendered hate speech*) : violence principalement contre les femmes et les filles en raison de leur identité de genre, intention de manquer de respect, d'introduire la peur et l'insécurité.

- Discours haineux religieux (*Religious hate speech*) : selon les pays, ce type de haine est surtout dirigé vers l'islam, l'hindouisme et le christianisme. Il est susceptible de nuire davantage puisqu'il peut atteindre un plus grand nombre de personnes.

- Discours haineux raciste (*Racist hate speech*) : discours ayant lieu au niveau international, la fréquence d'occurrence et l'impact du discours dépendant d'un pays à un autre.

- Discours haineux basé sur le handicap (*Hate speech on disability*) : discours de haine dû à la perception du handicap par le contrevenant, mais pas au handicap réel d'une personne.

- Discours de haine hybride (*Hybrid hate speech*) : discours de haine n'étant pas liée à un type particulier, haine pouvant être dirigée contre plus d'une communauté et d'une identité.

Force est de constater que les médias sociaux jouent un rôle majeur comme terreau fertile pour la prolifération des groupes racistes et d'extrême droite et leur propagande : « such websites [...] have become a key conduit through which extremists can educate others, transmit ideas and beliefs, and mobilise for demonstrations and rallies » (Banks 2010 : n.a.). Notre époque est en effet caractérisée par un accès relativement facile et peu couteux à Internet, ce qui accroit de façon importante la connectivité entre les personnes partout dans le monde. Dans ce contexte, Internet est devenu une « nouvelle frontière » de la propagation de la haine. Les réseaux sociaux, particulièrement, peuvent ainsi permettre de connecter des groupes qui étaient auparavant diversifiés et fragmentés. Cette

connectivité a pour conséquence de fédérer ces groupes, de générer une identité collective et de créer un sens de la communauté (Banks 2010; Gozlan 2013).

Les particularités de l'environnement en ligne peuvent ainsi jouer un rôle dans la propagation des discours haineux. En effet, Brown (2018) explique que certaines singularités d'Internet, à savoir la perception de l'anonymat, la distance (l'invisibilité) entre l'auteur d'un message et son récepteur, le sens de communauté et l'instantanéité des communications, jouent un rôle dans la distinction entre le discours de haine en ligne et le discours de haine hors ligne :

- L'anonymat (ou la perception de l'anonymat) : la perception d'une parole plus libre et sans censure a pour conséquence de supprimer la peur d'être tenu responsable des propos haineux que l'on peut propager en ligne.

- L'invisibilité : la distance physique entre l'émetteur et le récepteur d'un contenu – et vice versa – procure une certaine invisibilité qui fait en sorte que, d'une part, les effets immédiats du discours (par exemple la blessure émotionnelle) sont invisibles pour leur auteur et, d'autre part, les auteurs de haine en ligne ne sont pas en mesure de voir les visages des autres personnes qui désapprouvent leur contenu.

- La communauté : Internet réunit des gens séparés géographiquement, mais qui partagent les mêmes idées et qui peuvent entrer en contact les uns avec les autres. Cela peut accentuer les coalitions de personnes (*like-minded people*) au sein de groupes haineux visant à fédérer davantage de personnes pour partager leurs opinions.

- L'instantanéité : le délai entre une idée et son expression à une personne ou à un groupe de personnes peut s'articuler en quelques secondes seulement. Cela résulte ainsi en une certaine spontanéité des discours haineux due à cette instantanéité (manque de filtre ou réaction instinctive).

Ainsi, avec Internet, et particulièrement les médias sociaux, la capacité de diffuser la haine en ligne s'est accrue de façon considérable (Siapera et Viejo-Otero 2021). Ceci est d'autant plus problématique que la haine en ligne est liée, d'une part, à la cyberintimidation et au harcèlement en ligne et, d'autre part, à la cybercriminalité, dont certaines formes de harcèlement en ligne, de traque et de diffamation. En ce sens, le discours et les conduites haineux en ligne peuvent se concrétiser en d'autres types d'agressions quotidiennes (Keipi, Näsi, Oksanen et Räsänen 2016). À ce propos, Chetty et Alathur (2018) ont également montré le rôle que jouent les médias sociaux comme interface pour les discours haineux, les crimes haineux, l'extrémisme et le terrorisme (voir figure ci-dessous). En effet, alors que les discours haineux se propagent sur les réseaux sociaux par l'intermédiaire de publications et de messages des utilisateurs, ces plateformes peuvent également être utilisées comme espaces de coordination et de planification de crime haineux, aussi bien comme espaces de contact et de recrutement de personnes partageant

les mêmes idées par les groupes extrémistes et terroristes, soit pour diffuser la propagande, soit pour planifier et exécuter les attaques (Chetty et Alathur 2018).

Figure 1. Rôle central des médias sociaux pour les propos et conduites préjudiciables

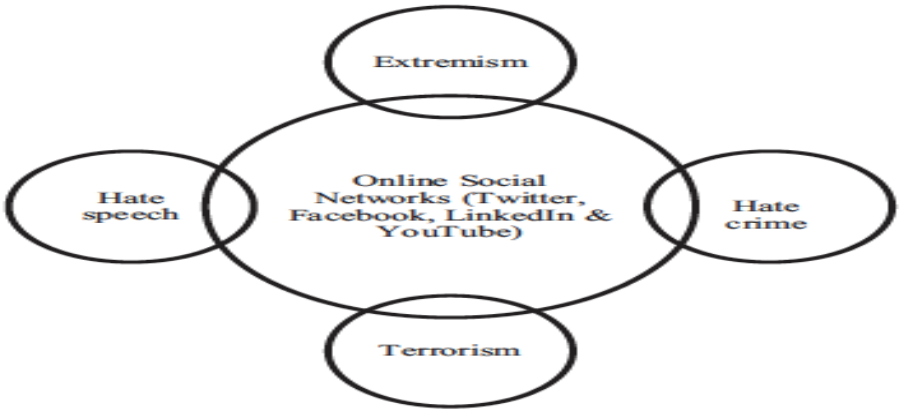

Fig. 1. Role of online social networks for destructive activities.

Source : Chetty et Alathur (2018).

La gouvernance des plateformes et les modèles de régulation

Le mode de gouvernance ou de régulation des plateformes dominant est appelé « autogouvernance » ou « autorégulation ». Selon Gorwa (2019b: 12-13), l'autogouvernance, comme approche, « limits platform liability and results in a relatively laissez-faire relationship between governing institutions and platform companies ». L'autogouvernance tend généralement à favoriser une transparence volontaire des plateformes, avec peu de supervision extérieure sur les processus de traitement ou de modération du contenu (Gorwa 2019b). La genèse de cette approche peut être associée au développement des sphères de sécurité (*safe harbor*) des plateformes en ligne aux États-Unis. En effet, le contexte de la prolifération de contenus illicites (notamment la pornographie) dans Internet dans les 1990 aux États-Unis a engendré des débats politiques et juridiques pour savoir à qui incombait la responsabilité de ces contenus illicites. Ainsi, en 1996, le Congrès votait la *Loi sur la décence dans les communications* (*Communication Decency Act*), qui faisait partie d'une loi plus vaste sur les télécommunications (*Telecommunications Act of 1996*) pour faire face à ces enjeux. Cette loi contient un article particulier qui a transformé de façon historique le régime de neutralité des plateformes d'intermédiation en ligne : la Section 230 (Gorwa 2019a et 2019b; Siapera et Viejo-Otero 2020; Gilespie 2010, 2017, 2018a et 2018b). Cette disposition offre des sphères de sécurité contre toute responsabilité pour les contenus préjudiciables que leurs utilisateurs pourraient fournir en ligne (Gillespie 2017, 2018a et 2018b). La pratique des *safe harbors*

s'est diffusée à travers le monde et notamment en Europe où la Directive sur le commerce électronique (*E-Commerce Directive*) de la CE fixe le cadre législatif de la responsabilité des plateformes ou intermédiaires en ligne. Celle-ci stipule que les « fournisseurs de services d'intermédiation » sont exemptés de responsabilité pour les contenus qu'ils stockent ou hébergent, et ce, s'ils agissent de manière strictement passive » (*E-Commerce Directive* 2020).

Gorwa (2019b) soutient que l'autogouvernance peut avoir quelques avantages. D'abord, la pression publique, venant notamment du journalisme d'investigation, du milieu académique ou des groupes de la société civile, peut pousser les entreprises à entreprendre des changements sans pour autant une intervention législative complexe. De même, la délégation de la prise de décision sur la liberté d'expression (par exemple via la modération) aux intermédiaires en ligne apaiserait les craintes relatives à la censure et à la répression du gouvernement (surtout dans les pays moins démocratiques). Or, ce laisser-faire via des accords volontaires repose justement sur la bonne volonté de ces acteurs, ce qui offre peu de possibilités de recours en cas de non-respect des obligations. Les initiatives de transparence mises en place par les intermédiaires en ligne sont davantage tournées vers le public et offrent peu d'informations utiles aux législateurs et aux régulateurs (Gorwa 2019b).

Ainsi, le mode de gouvernance des plateformes relève d'une importance capitale, dans la mesure où les décisions que ces acteurs prennent façonnent non seulement les communications entre les gens, mais aussi la vie sociale et politique des utilisateurs (Suzor 2018). En effet, les médias sociaux jouent un rôle important dans la promotion et la limitation de la liberté d'expression en ligne (DeNardis et Hackl 2015). En ce sens, ces plateformes ont recours à un outil indispensable au cœur de leur fonction : la modération (Gillespie 2010, 2017, 2018a et 2018b). La modération est un processus utilisé par ces plateformes pour déterminer si les contenus préjudiciables doivent être supprimés ou non (Ullmann et Tomalin 2020). Les signalements sont alors des outils pratiques et symboliques permettant aux plateformes de maintenir un niveau d'autorégulation et, ainsi, d'éviter l'intervention gouvernementale sur les mécanismes de traitement de contenus (Crawford et Gillespie 2016).

Il existe une distinction entre la modération de contenu réactive et la modération proactive (Bakalis et Hornle 2021). La modération réactive implique principalement le signalement de contenus par les utilisateurs et le retrait de celui-ci ou le blocage d'un compte ou d'un groupe par les plateformes, selon la gravité du contenu. C'est ce que l'on qualifie de modèle *publish-then-filter* (Gillespie 2017 : 16), que l'on peut traduire par « publier puis filtrer ».

> This means that even heinous content may get published, at least briefly, and criminal behavior may occur (and have its intended impact) before anything is done in response. Plenty of content that violates site guidelines remains online for days, or years, because of the sheer challenge of policing platforms as immense as these (Gillespie 2017 : 16).

La modération proactive, quant à elle, implique plutôt la prévention proactive, permettant de bloquer le contenu lors de sa publication et avant sa notification par un utilisateur

(Bakalis et Hornle 2021; Gillespie 2010 2018). Cette forme de modération s'exerce notamment grâce à la technologie. Pour les législateurs, il y a une tentation d'appeler à l'automatisation de la modération grâce à l'utilisation de l'intelligence artificielle (Bakalis et Hornle 2021).

Toutefois, ces deux modèles de modération ne sont pas sans défauts. En effet, la modération réactive comprend plusieurs défis pour prévenir et lutter contre les discours haineux en ligne, étant donné que le processus de notification et de retrait de contenu est plus lent, et ce, même si la plateforme décide de retirer le contenu après des heures (Bakalis et Hornle 2021). Ce processus pose également un défi juridique, dans la mesure où il est impossible pour les plateformes de garantir qu'aucun contenu ou comportement illégal n'y apparaitra. En outre, il y a un défi éthique, puisque les utilisateurs ne sont pas à l'abri de l'obscénité et ne peuvent être totalement protégés du contenu préjudiciable (Gillespie 2017). La modération proactive représente également un défi de taille pour les grandes plateformes, étant donné le volume de contenus et d'activités qui sont publiés, où le rôle du modérateur consisterait à évaluer chaque contenu avant son apparition sur le site (Gillespie 2017). Bref, la modération est un processus complexe, qui nécessite d'immenses ressources humaines (Gillespie 2018a et 2018b).

Le second modèle de régulation implique davantage d'intervention étatique. En ce sens, les plateformes sont soumises aux normes et lois nationales des pays où elles opèrent. Elles peuvent ainsi être redevables aux exigences des autorités étatiques que ce soit en matière de transparence et de divulgation d'information, de demande de suppression d'accès aux informations répréhensibles ou encore en matière de conformité des technologies afin de faciliter l'application des lois nationales (Gillespie 2017, 2018a et 2018b; Gorwa 2019a et 2019b; Suzor 2018).

En matière de régulation des plateformes et des contenus en ligne, les différents acteurs étatiques mettent en œuvre des « rationalités de gouvernement et des stratégies de régulation » (Badouard 2021 : 89) qui diffèrent selon un continuum de contrainte. On parle alors de *soft governance*, qui consiste en la mise en place de normes non contraignantes censées tout de même produire des résultats, ou de *hard governance*, qui découle des normes plus strictes, par exemple les lois, les règlements, les directives ou les traités (Maggetti 2015). Comme le propose Badouard (2021), ces approches peuvent être adaptées à notre analyse des tentatives de régulation des plateformes en ligne. Nous parlerons alors de *hard regulation* et de *soft regulation*.

L'approche *hard regulation* implique traditionnellement un acteur de régulation, en l'occurrence l'État, qui édicte des règles et des normes sans collaboration directe d'autres parties prenantes (Gorwa 2019a et 2019b). Selon Badouard (2021 : 112), l'approche *hard regulation* 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 ». En matière de lutte contre les discours de haine en ligne, l'Allemagne a opté pour cette approche. De l'autre côté, l'approche de *soft regulation* est davantage basée sur une collaboration entre l'acteur étatique et les parties prenantes, en l'occurrence les acteurs sujets de régulation. En effet, l'approche de *soft regulation* « consiste en la mise en place d'un partenariat avec les plate-

formes 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 » (Badouard 2021 : 89-90).

Cette forme de régulation implique souvent des codes volontaires de conduite et des accords d'autorégulation. On la qualifie aussi de « corégulation » (Gorwa 2019a et 2019b). La corégulation ou co-gouvernance se situe en fait entre l'intervention étatique (gouvernance DES plateformes) et l'autogouvernance (gouvernance PAR les plateformes). L'Union européenne est considérée comme un chef de file en la matière (Gorwa 2019a et 2019b; Quintel et Ulrich 2020). On peut penser au Code européen de bonnes pratiques contre la désinformation (Commission européenne 2018) ou au Code de conduite pour lutter contre les discours de haine illégaux en ligne (Commission européenne 2016).

Éléments méthodologiques

Approche de recherche

Le choix des juridictions ou acteurs étatiques à l'étude dans ce rapport se base sur des critères d'inclusion que nous avons jugés adéquats par rapport à l'objet de recherche. Selon ces critères, le cadre réglementaire doit : 1) cibler les plateformes en ligne, y compris les plateformes de médias sociaux; 2) viser les préjudices en ligne, y compris les discours haineux; 3) être articulé dans une loi ou un Livre blanc; 4) être pleinement en vigueur (par exemple ne pas avoir été invalidé par un organe constitutionnel) et 5) avoir produit des résultats mesurables. À la lumière de ces critères, notre analyse portera sur deux cadres réglementaires : la *Loi sur l'application des réseaux* ou NetzDG de l'Allemagne et le *Code de conduite visant à combattre les discours de haine illégaux en ligne* de la CE.

Pour rappel, ce document poursuit deux objectifs : dans un premier temps, il vise à décrire les approches de régulation, de même que les instruments de politique (*policy tools*) mis en place par ces cadres réglementaires; dans un second temps, une analyse comparative de ces deux cadres réglementaires sera fournie sur les bases d'une typologie des approches de régulation et d'instruments de politiques, mais aussi de conformité des acteurs visés aux différentes obligations législatives par le biais de certains indicateurs.

Pour ce faire, le choix de la méthodologie de recherche a été porté sur l'étude de cas multiples. Selon Gagnon et Fortin (2016 : 197), « [l]'étude de cas consiste à faire état d'une situation réelle particulière, prise dans son contexte, et à l'analyser pour découvrir comment se manifestent et évoluent les phénomènes auxquels le chercheur s'intéresse ». Les travaux de Robert Yin et de Robert Stake sont souvent cités comme ayant contribué de façon significative à la reconnaissance de l'étude de cas dans l'explication des « liens complexes d'un phénomène contemporain dans son contexte de vie réelle » (Alexandre, 2013 : 28). L'étude de cas permet d'analyser en profondeur et de comparer plusieurs cas sur la base d'une démarche objective. Robert Yin propose une vision positiviste de l'étude de cas, qu'il conçoit comme une enquête empirique permettant de tester et de corroborer une hypothèse. Robert Yin et Robert Stake mettent ainsi de l'avant une démarche analy-

tique déductive des études de cas pour permettre au chercheur une reproductibilité des résultats issus de la recherche (cité dans Alexandre 2013).

Technique de collecte de données

L'étude de cas permet l'utilisation d'une diversité de sources de preuves ou de données « de manière à décrire en profondeur le phénomène étudié » (Gagnon et Fortin 2016 : 199). Parmi les méthodes de collecte de données qualitatives (observation, entrevue, groupe de discussion, incident critique, etc.), nous avons opté pour l'analyse documentaire, qui nous parait la technique de collecte la plus adéquate. En effet, la nature de notre analyse nous amène à nous pencher sur différents types de documents : projets de loi, documents de politique (*policy paper*), rapports, articles de journaux, enquêtes publiques, articles scientifiques, etc.

Concrètement, cette analyse portera d'abord sur les textes édictant les cadres réglementaires, à savoir les textes de loi (*Network Enforcement Act* ou NetzDG dans le cas de l'Allemagne et les codes de conduite (dans le cas de la Commission européenne). À ces textes s'ajoute un ensemble de documents complémentaires qui comprend les lignes directrices (*guidelines*), les rapports gouvernementaux ou des agences de régulation, les évaluations des politiques effectuées par les organes gouvernementaux ou encore les feuilles de route (*roadmap*) publiés par les gouvernements. En second lieu seront analysés les documents produits par les plateformes en ligne pour répondre aux contraintes législatives : rapports de transparence, règles de communautés, articles en ligne ou toute communication publique relative à la législation les visant.

Les hypothèses de recherche

Comme il a déjà été indiqué, l'approche positiviste de Yin propose que l'étude de cas soit conçue comme une enquête empirique permettant de valider des hypothèses émises par le chercheur. Notre modeste démarche de recherche ne prétend pas remplir une telle exigence scientifique. Néanmoins, l'analyse comparative d'approches de régulation au cœur de ce document entend tirer des leçons et des conclusions de ces expériences vécues ailleurs. À cet effet, une précision est de mise : le but du présent document ne s'inscrit pas dans une démarche d'évaluation de politiques publiques ou de programmes – ce qui n'est d'ailleurs pas notre spécialisation. Ainsi, la recherche ne comprendra pas une définition des questions de l'évaluation, encore moins un choix des critères d'appréciation (efficacité, efficience, pertinence, etc.) ni des stratégies de valorisation des résultats de l'analyse (Knoepfel, Larrue, Varone et Savard 2015).

La particularité de notre démarche d'analyse réside toutefois dans le fait que les cadres de réglementation à l'étude, bien que poursuivant les mêmes fins, utilisent des moyens relativement différents pour y parvenir. Nous proposons tout de même quatre hypothèses centrales à vérifier lors de l'analyse de ces approches de régulation. Nous voulons particulièrement vérifier dans quelle mesure chaque approche répond à ces hypothèses :

- Les mesures contraignantes relatives à la modération des contenus dictées par la loi allemande ont résulté en blocages ou suppressions disproportionnés de contenus par les plateformes;

- Les obligations en vertu d'une approche de régulation douce (*soft regulation*) sont moins respectées par les signataires que dans le cadre d'obligations contraignantes (*hard regulation*);

- Globalement, les mesures de régulation ont favorisé davantage de transparence sur les pratiques de modération de contenus de la part des plateformes;

- Globalement, les mesures ont permis une diminution importante du nombre de publications répréhensibles ou préjudiciables signalées et annuellement rapportées sur les plateformes.

Résultats

Les deux approches de régulation analysées dans ce chapitre sont conçues en fonction du niveau de contraintes qu'elles imposent aux plateformes en ligne. Pour rappel, la réglementation stricte (*hard regulation*) se fait notamment à travers des instruments obligatoires (par exemple des lois), alors que la réglementation douce (*soft regulation*) se fait à partir d'instruments volontaires (par exemple des codes de conduite) (Rapp, Schmid et Wolff 2011).

La loi allemande *Netzwerkdurchsetzungsgesetz* (ou NetzDG), entrée en vigueur le 1er janvier 2018, vise principalement l'amélioration du traitement des contenus illégaux par les réseaux sociaux (*Netzwerkdurchsetzungsgesetz* 2018). Ainsi, l'approche allemande peut être associée à la *hard regulation*, qui consiste à « imposer aux plateformes des obligations de résultats et/ou de délais à respecter pour le retrait des contenus problématiques, et elle prévoit [sic] des sanctions en cas de non-mise en conformité avec les nouvelles règles » (Badouard 2021 : 89).

Le Code de conduite de l'UE pour lutter contre les discours de haine illégaux en ligne, mis en place en 2016 en collaboration avec quatre plateformes majeures (Facebook, Microsoft, Twitter et YouTube), s'articule dans la longue tradition d'application de la corégulation ou de l'autorégulation privilégiée par l'Union européenne (Banks 2019; Gorwa 2019a et 2019b; Quintel et Ullrich 2020). Dans le cadre de notre analyse, nous préférons utiliser le concept de *soft regulation* (Rapp, Schmid et Wolff 2011; Badouard 2021). En l'occurrence, la *soft regulation* est une approche qui repose 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 » (Badouard 2021 : 113). Elle permet ainsi aux parties de mettre en place des obligations de moyens plutôt que des obligations de résultat.

Portée de la régulation

La loi NetzDG cible notamment 21 incriminations définies dans le Code pénal allemand et pouvant faire l'objet d'un retrait, dont, entre autres, l'insulte et la diffamation, l'incitation à la haine ou au crime, les menaces de mort, la diffusion de matériel de propagande, l'utilisation de symboles d'organisations inconstitutionnelles, etc. La loi NetzDG

s'applique aux fournisseurs de réseaux sociaux ayant plus de deux (2) millions d'utilisateurs enregistrés en République fédérale d'Allemagne. Cela inclut, par exemple, Facebook, Instagram, Twitter et YouTube. Les plateformes de messagerie privée telles que WhatsApp et Telegram sont exemptés de l'application de la loi. En ce sens, cette dernière cible principalement les communications publiques, de même que les plateformes ayant pour fonction principale de permettre aux utilisateurs de partager et de communiquer du contenu avec d'autres d'utilisateurs.

Le Code de conduite de la Commission européenne, quant à lui, vise à prévenir et à lutter contre les discours de haine illégaux en ligne, notamment en s'appuyant sur une collaboration avec quatre plateformes majeures, à savoir Facebook, Microsoft, Twitter et YouTube. En plus des signataires de départ, cinq (5) autres plateformes majeures se sont également jointes au Code de conduite, notamment Instagram (2018), Snapchat (2018), Dailymotion (2018), Jeuxvideo.com (2019) et TikTok (2020). Par ailleurs, le Code de conduite est construit sur la *Décision-cadre 2008/913/JAI du Conseil de l'Union européenne sur la lutte contre certaines formes et manifestations de racisme et de xénophobie au moyen du droit pénal* (Décision-cadre 2008/913/JAI), qui définit le discours haineux comme toute « incitation publique à la violence ou à la haine visant un groupe de personnes ou un membre d'un tel groupe, défini par référence à la race, la couleur, la religion, l'ascendance, l'origine nationale ou ethnique » (art.1(a)).

Survol des principales obligations

Sous la loi NetzDG, les plateformes doivent mettre en place un mécanisme de signalement de « contenus pénalement punissables » qui soit facile d'accès et disponible en tout temps. Tout contenu « manifestement illégal » doit être supprimé ou bloqué dans les 24 heures suivant la réception d'une plainte jugée fondée. Pour tout contenu dont l'illégalité est difficile à déterminer, les plateformes disposent de sept jours pour le supprimer ou le bloquer. Elles peuvent également renvoyer la décision à une institution d'autorégulation reconnue, qui doit alors se prononcer sur l'illégalité du contenu dans un délai de sept jours. Enfin, les plateformes doivent communiquer au plaignant et à l'auteur du contenu toutes les décisions prises concernant la plainte et le raisonnement derrière son acceptation ou son rejet.

De plus, les plateformes recevant plus de 100 plaintes par année civile concernant des contenus illicites ont l'obligation de rendre publics des rapports semestriels en allemand sur la gestion des plaintes. Ceux-ci doivent contenir des informations entre autres sur le volume de plaintes reçues et la manière dont elles ont été traitées. Les plateformes doivent nommer un représentant en Allemagne qui soit responsable de la réception des demandes que les autorités allemandes pourraient leur adresser. En outre, la loi NetzDG institue un *droit à la divulgation* (*right to disclosure*), permettant à un plaignant d'exiger que le réseau social concerné divulgue les informations sur l'auteur de l'infraction, et ce, sous l'ordonnance d'un tribunal civil compétent.

La loi NetzDG prévoit également un mécanisme coercitif. Les lignes directrices sur la fixation des amendes publiées par le ministère de la Justice pour les violations de la NetzDG stipulent que l'amende maximale pour le non-respect des obligations est fixée à 40 millions d'euros. En outre, le niveau des amendes dépend de la taille du réseau social

et de la gravité de la violation de la loi. Les lignes directrices divisent les réseaux sociaux en trois catégories : les réseaux sociaux de catégorie A – comptant plus de 20 millions d'utilisateurs enregistrés (par exemple Facebook) – s'exposent à une amende maximale de 40 millions d'euros. Les réseaux sociaux de catégorie B – comptant de 4 à 20 millions d'utilisateurs enregistrés (par exemple Instagram et YouTube) – s'exposent à une amende maximale de 25 millions d'euros. Enfin, les réseaux sociaux de catégorie C – comptant deux à quatre millions d'utilisateurs enregistrés (par exemple Twitter) – peuvent être condamnés à une amende maximale de 15 millions d'euros. Ces lignes directrices précisent également que des amendes ne seront infligées aux réseaux sociaux qu'en cas de non-traitement systémique des plaintes. Autrement dit, aucune amende n'est prévue dans les cas isolés ou de violations ponctuelles de la loi. En outre, le ministère de la Justice doit obtenir une ordonnance de la Cour pour infliger une amende à un réseau social en cas de manquement à l'obligation de supprimer ou de bloquer tout contenu illégal.

Le Code de conduite de la CE établit des obligations non contraignantes auxquelles les plateformes signataires s'engagent. On peut citer en exemple la mise en place de règles et de normes de communautés (*community guidelines*) interdisant les discours de haine, la mise en place de systèmes (algorithmes) et d'équipes pour examiner le contenu qui contrevient à ces normes, l'examen de la majorité du contenu signalé dans un délai de 24 heures et la suppression ou le blocage de l'accès au contenu haineux, la formation régulière des modérateurs, l'engagement dans des partenariats et des activités de formation avec les organisations de la société civile (OSC) dans le but d'élargir le réseau de « signaleurs de confiance » (*trusted flaggers*) et la promotion de la transparence vis-à-vis des utilisateurs et du grand public.

Le Code de conduite est donc en quelque sorte un contrat entre la CE et les géants numériques, qui s'y engagent de plein gré. En ce sens, il n'est pas juridiquement contraignant et n'inclut pas de dispositions relatives à l'application (*enforcement*), aux sanctions ou à la conformité en dehors des lois en vigueur dans les États membres. En revanche, les compagnies signataires ont consenti à ce que la CE évalue le niveau de respect de leurs engagements au Code de conduite, en plus de leur impact.

Les rapports de transparence sous la NetzDG

Les rapports de transparence fournis par les médias sociaux facilitent en partie l'appréciation de leur conformité à la régulation allemande. Nous focaliserons notre analyse sur les plus grandes plateformes, à savoir les plateformes de catégories catégorie A et B : Facebook, Instagram et YouTube. Les rapports de transparence de ces acteurs nous fournissent beaucoup d'informations qualitatives et quantitatives permettant de catégoriser en quelque sorte leurs pratiques.

Mécanismes de signalement des plaintes

Pour se conformer à la législation allemande, les plateformes en ligne, notamment Facebook, Instagram et YouTube, ont mis en place un mécanisme de traitement de contenu à deux échelles : dans un premier temps, le contenu est d'abord traité selon leurs normes de communauté (*Community Standards* chez Facebook et Instagram, *Community Guidelines* chez Google/YouTube), qui définissent ce qui est autorisé ou interdit sur leurs plate-

formes; dans un second temps, les modérateurs jugent de la conformité ou de la légalité du contenu selon le droit allemand (Badouard 2020; 2021; Wagner, Rozgonyi, Sekwenz, Cobbe et Singh 2020; Instagram 2019, 2020a, 2020b et 2021; Facebook 2018, 2019a, 2019b, 2020a, 2020b et 2021; YouTube 2018, 2019a, 2019b, 2020a, 2020b et 2021). Dans le cas de Facebook et de sa plateforme Instagram, la compagnie a mis en place un formulaire de signalement NetzDG (*NetzDG reporting form*), exclusivement accessible aux citoyens allemands, qui permet à ceux-ci de signaler tout contenu qu'ils jugent enfreindre les dispositions du Code pénal allemand énoncées dans la loi NetzDG. YouTube a quant à lui mis en place deux mécanismes de signalement : le premier, intégré dans le flux de signalement, est accessible sous chaque vidéo et à côté de chaque commentaire; le second est un formulaire légal Web également accessible dans le menu principal de YouTube.

Ressources humaines et organisationnelles

Les plateformes ont également mis en place des ressources organisationnelles et humaines pour assurer la conformité à la législation. À titre d'illustration, chez Facebook et Instagram, les contenus évalués selon les normes de la compagnie sont traités par une équipe chargée des opérations du marché mondial (*Global Market Operations team*), alors que les contenus évalués à l'étape de la légalité selon la loi NetzDG sont traités par une équipe d'avocats (Legal Takedown Request Operations team). Selon Facebook, cette procédure permet de s'assurer que les contenus manifestement illégaux soient retirés ou bloqués dans les 24 heures. Dans les cas où il est difficile de se prononcer sur la légalité du contenu, Facebook se dirige alors vers le FSM (Freiwillige Selbstkontrolle Multimedia-Diensteanbieter e.V.), une autorité d'autorégulation allemande pour une consultation juridique externe (Instagram 2019, 2020a, 2020b et 2021; Facebook 2018, 2019a, 2019b 2020a, 2020b et 2021). YouTube, de son côté, utilise un processus similaire sur sa plateforme: alors que les contenus évalués selon les *Community Guideline* sont traités par les équipes d'examen composées de plusieurs germanophones, les requêtes relatives à la loi NetzDG sont traitées par des avocats (NetzDG team) (YouTube 2018, 2019a, 2019b, 2020a, 2020b et 2021).

Correspondance et transparence avec les utilisateurs

Généralement, les plateformes vont communiquer dans le cadre du processus de signalement avec le plaignant. L'information communiquée peut prendre plusieurs formes selon la décision qui résulte de l'examen de la plainte : soit l'on communique effectivement la violation des normes de communauté, soit l'on justifie le retrait en vertu de la violation des dispositions du Code pénal allemand. Dans d'autres circonstances, l'on communique au plaignant que le contenu ne viole ni les normes de communauté ni le Code pénal allemand.

Analyse de principaux indicateurs

Nous avons compilé (sur Excel) les données — grâce aux informations fournies par les rapports de transparence — sur le traitement des plaintes réalisé entre 2018 et 2021 par les plateformes. Ces données se basent sur certains des indicateurs visant à mesurer l'évolution dans le temps sur Facebook (tableau 1), sur Instagram (tableau 2) et sur YouTube (tableau 3). Les principaux indicateurs sont le volume de signalement reçus (Face-

book et Instagram) ou le nombre d'éléments (vidéos et commentaires) signalés (YouTube), le volume de signalement ayant entrainé le blocage ou la suppression de contenus (Facebook et Instagram), le volume de contenus répréhensibles supprimés et le volume de contenus supprimés en 24 heures.

Une des principales raisons des différences de dénomination des indicateurs réside dans le fait que sur Facebook et sur Instagram, une plainte peut concerner plusieurs infractions, ce qui justifie la différence entre le volume de plaintes reçues et celui des infractions signalées (par exemple, sur Facebook, 886 plaintes ont signalé 1704 infractions criminelles en 2018).

Tableau 1. Traitement des signalements sur Facebook

Indicateurs	Évolution par année											
	Juillet 2018		Janvier 2019		Juillet 2019		Janvier 2020		Juillet 2020		Janvier 2021	
	Nb.	%	Nb.	%	Nb.	%	Nb.	%	Nb.	%	Nb.	%
Nb. formulaires de signalement NetzGD reçus	886		500		674		3087		4292		4211	
Nb. de contenus répréhensibles signalés	1704		1048		1050		4274		6038		4401	
Volume de plaintes ayant entraîné la suppression ou le blocage de contenu	218	24.60	159	31.80	239	35.46	562	18.21	1344	31.31	1117	26.53
Volume de contenus répréhensibles supprimés	362	21.24	369	35.21	349	33.24	1043	24.40	2308	38.22	1276	28.99
Volume de plaintes ayant mené à la suppression ou au blocage de contenu en 24h			108	67.92	204	85.36	488	86.83	1229	91.44	1013	90.69
Volume de plaintes pour lesquelles un conseiller juridique externe a été consulté	54	6.09	23	4.6	13	1.93	14	0.45	8	0.19	6	0.14

Source : Rapports de transparence NetzDG, Facebook (2018, 2019a, 2019b, 2020a, 2020b et 2021).

Tableau 2. Traitement des signalements sur Instagram

Indicateurs	Évolution par année							
	Juillet 2019		Janvier 2020		Juillet 2020		Janvier 2021	
	Nb.	%	Nb.	%	Nb.	%	Nb.	%
Nb. formulaires de signalement NetzGD reçus	146		365		2025		3366	
Nb. de contenus répréhensibles signalés	252		468		3458		5570	
Volume de plaintes ayant entraîné la suppression ou le blocage de contenu	38	26.03	169	46.30	694	34.27	415	12.33
Volume de contenus répréhensibles ou préjudiciables supprimés	116	46.03	221	47.22	1067	30.86	884	15.87
Volume de contenus répréhensibles supprimés en 24h	23	19.83	92	41.63	595	55.76	242	27.38
Volume de plaintes pour lesquelles un conseiller juridique externe a été consulté	2	1.37	2	0.55	3	0.15	2	0.06

Source : Rapports de transparence NetzDG, Instagram (2019, 2020a, 2020b et 2021).

Tableau 3. Traitement des signalements sur YouTube

Indicateurs	Évolution par année											
	Juillet 2018		Janvier 2019		Juillet 2019		Janvier 2020		Juillet 2020		Janvier 2021	
	Nb.	%	Nb.	%	Nb.	%	Nb.	%	Nb.	%	Nb.	%
Éléments (vidéo ou commentaire) signalés	214,827		250,957		304,425		277,478		388,824		323,792	
Éléments (vidéo ou commentaire) supprimés ou bloqués	58,297	27.14	54,644	21.77	71,168	23.38	71,907	25.91	90,814	23.4	73,477	22.69
Éléments (vidéo ou commentaire) signalés mais non supprimés ou bloqués	156,530	72.86	196,313	78.23	233,257	76.62	205,571	74.09	298,010	76.6	250,315	77.31
Délai de suppression/Volume d'éléments supprimés en 24h	54,199	92.97	52,030	95.22	62,492	87.81	66,309	92.21	83,706	92.17	64,774	88.16

Source : Rapports de transparence, YouTube (2018, 2019a, 2019b, 2010a, 2020b et 2021).

Ces données permettent de constater une constance dans le taux de suppression de contenus qui, chez Facebook, par exemple, n'a atteint que 38,2 % dans le rapport de juillet 2020. Cela représente seulement 31,3 % des plaintes ayant mené à cette suppression. Chez Instagram, ce taux de suppression n'a cessé de baisser depuis 2019, passant de 46 % à 15,9 % dans le rapport de janvier 2021, ce qui représentait 26 % des plaintes en 2019, contre 12,3 % en 2021. Toutefois, si Facebook affiche un taux de contenu supprimé en 24 heures relativement encourageant (plus de 90 % dans le premier rapport semestriel de 2021), l'on ne peut pas en dire autant d'Instagram. Chez YouTube, la proportion de contenus supprimés est relativement faible : passant de 27,1 % en 2018 à 22,7 % seulement dans le premier rapport semestriel de 2021. À l'inverse, la proportion de contenus supprimés en 24 heures semble baisser, ayant enregistré la plus forte hausse dans le premier rapport semestriel de 2019 (95,2 %). Ce taux se situe à 88,2 % dans le dernier rapport, loin devant celui d'Instagram pour le même semestre.

Il est également important de mentionner que l'augmentation du nombre de plaintes ou de contenus signalés – jugés répréhensibles – ne se conjugue pas forcément avec les taux de suppression de contenus, qui sont demeurés stables au fil du temps. En effet, les données exposées ci-haut indiquent que moins de la moitié des plaintes ou de contenus signalés conduisent à une suppression, même si le nombre de signalement a continué d'augmenter. Néanmoins, il est sans doute possible d'attribuer cette augmentation de signalements au fait que les gens sont plus au courant des recours possibles et dénoncent davantage, sans pour autant qu'il n'y ait une croissance de contenus jugés répréhensibles par les plateformes en ligne.

Un autre indicateur que nous avons analysé concerne la consultation extérieure pour juger de l'illégalité d'un contenu, qui se fait au moyen d'un conseiller juridique externe ou d'une institution d'autorégulation. Pour rappel, les contenus qui ne sont pas manifestement illégaux, c'est-à-dire dont l'illégalité peut être difficile à déterminer, peuvent faire l'objet d'un examen externe. Ainsi, les plateformes disposent de sept jours pour statuer sur une telle illégalité. Si la consultation de conseillers juridiques est relativement faible (118 chez Facebook, neuf chez Instagram et 20 chez YouTube[4]), YouTube rapporte n'avoir consulté l'institution d'autorégulation que pour huit éléments signalés. Jusqu'ici,

[4] Les rapports de YouTube ne permettent pas de ventiler l'indicateur par année, d'où son absence dans le tableau.

Facebook et Instagram n'ont envoyé respectivement qu'un seul signalement à cet organisme.

La mise en œuvre des engagements des plateformes signataires

La CE évalue la mise en œuvre du Code de conduite par les plateformes signataires au moyen d'un exercice régulier mis en place en collaboration avec des organisations de la société civile de différents pays de l'UE. L'évaluation se penche sur plusieurs indicateurs, dont 1) le nombre de signalements de contenus haineux reçus chaque année; 2) le taux de suppression, c'est-à-dire le taux de contenus considérés comme des discours de haine illégaux effectivement supprimés après le traitement du signalement; 3) le volume de contenus manifestement illégaux signalés aux autorités policières; 4) le délai de traitement des signalements (en 24 heures, en 48 heures, etc.); 5) le taux de plaintes pour lesquelles les plateformes effectuent un retour d'information aux utilisateurs et 6) les motifs de dénonciation de la haine (xénophobie, orientation sexuelle, identité de genre, religion, etc.) (Commission européenne 2016, 2017, 2018, 2019 et 2020).

En septembre 2019, la CE a publié une évaluation globale de la mise en œuvre des engagements des plateformes signataires entre 2016 et 2019. Ses conclusions font valoir que, globalement, le Code de conduite a été efficace et a abouti à des progrès rapides, notamment en ce qui a trait à l'examen et à la suppression rapides des discours de haine. À titre d'illustration, la CE constate que 28 % du contenu signalé a été supprimé en 2016 contre 72 % en 2019 (indicateur 2); 40 % des avis de signalements ont été examinés dans les 24 heures en 2016 contre 89 % en 2019 (indicateur 4). La CE fait également valoir que le Code a renforcé la confiance et la coopération entre les plateformes, les organisations de la société civile et les autorités des États membres sous la forme d'un processus structuré d'apprentissage mutuel et d'échange de connaissances. Toutefois, la CE fait remarquer qu'il demeure des améliorations à apporter en termes de transparence et de retour d'informations aux utilisateurs (indicateur 5), notant que le retour d'informations aux utilisateurs fait toujours défaut pour près d'un tiers des notifications en moyenne. Enfin, la CE déplore le peu de collaboration entre les plateformes et les autorités policières, notamment le faible volume de contenus manifestement illégaux signalés aux autorités compétentes (indicateur 3) (Commission européenne 2019).

Pour nous assurer de comprendre l'évolution de ces indicateurs dans le temps, notamment entre 2016 et 2020, de même que pour juger de l'efficacité de la réponse des plateformes dans le traitement des signalements, nous avons réalisé une compilation de données à partir des informations fournies par les rapports de la CE. Le tableau 4 offre un portrait global de l'évolution des indicateurs 1, 2, 3, 4, 5[5].

[5] Les motifs de dénonciation ou de signalement ne permettent pas en soi de témoigner de l'efficacité des dispositifs mis en place par les plateformes.

Tableau 4. Évolution des indicateurs de signalements entre 2016 et 2020

Indicateurs	Évolution par année				
	2016	2017	2018	2019	2020
Nb. de signalements de contenus de haine illégaux reçus par les plateformes	600	2,575	2,982	4,392	4,364
Volume (en %) de contenus considérés comme discours de haine illégaux supprimés	28.2%	59.1%	70%	71.7%	71%
N.b de contenus manifestement illégaux signalés aux autorités policières	-	212	511	503	475
Volume (en %) de signalements évalués par les plateformes dans un délai de 24 h	40%	51.40%	81.7%	88.9%	90,4%
Volume (en %) des notifications reçues ayant une réponse et un retour d'information des plateformes	-	-	68.90%	65.40%	67.10%

Source : Commission européenne (2016, 2017, 2018, 2019 et 2020)

Lorsqu'on analyse les données sur le respect des engagements des plateformes relatifs au Code de conduite, il est possible de remarquer que les résultats du premier *monitoring* de la mise en œuvre sont négatifs. L'analyse de l'évolution des indicateurs permet toutefois de constater que, sur une longue période, l'application des engagements s'est nettement améliorée. Par exemple, l'on constate une amélioration dans le temps du délai de traitement de contenus haineux signalés, de même que de la proportion de contenus effectivement illégaux, par exemple entre 2018 et 2020. En revanche, le nombre de contenus illégaux signalés aux autorités policières est plutôt faible, sachant que 71 % de contenus supprimés en 2020 représentent environ 3099 de contenus illégaux (sur 4364). En ce sens, 475 cas signalés aux autorités policières ne représentent que 15,3 % de contenus illégaux supprimés. En matière de transparence, moins de sept plaintes font l'objet d'une correspondance avec les utilisateurs. Si certaines de ces données peuvent paraitre encourageantes, il est néanmoins difficile de juger de leur réelle efficacité en raison du manque d'accès direct aux rapports des plateformes. En effet, les rapports de *monitoring* fournis par la Commission européenne sont en fait des fiches d'information de quatre à cinq pages en moyenne.

Analyse comparative des approches de régulation

Les deux approches de régulation analysées poursuivent les mêmes objectifs, mais empruntent des méthodes relativement différentes (voir la schématisation dans le tableau 5). La loi NetzDG couvre un large éventail de préjudices en ligne, qui constituent 21 contenus et comportements étant définis comme des infractions au Code criminel allemand. En ce sens, l'illégalité du contenu clairement défini dans la loi facilite sans doute l'examen de celui-ci par les plateformes en ligne. En réalité, la loi NetzDG vise l'application et le renforcement par les médias sociaux de règles déjà existantes. Cela est d'autant plus vrai que l'obligation de suppression de contenu n'est pas nouvelle; en effet, en vertu de la *Loi sur les Télémédias*, qui transpose la directive européenne sur le commerce électronique (E-Commerce Directive) en Allemagne, les plateformes de médias sociaux sont tenues responsables lorsqu'elles ont connaissance d'un contenu illégal et qu'elles ne sont pas parvenues à le supprimer sans retard injustifié (*Netzwerkdurchsetzungsgesetz* 2018; Theil 2019). En revanche, la portée du Code de conduite européen est beaucoup moins grande et ne s'applique qu'au discours haineux, tel que défini dans la Décision-cadre 2008/913/JAI. Comme la loi NetzDG, le Code vise principalement la pré-

vention et la lutte contre les discours de haine déjà illégaux. Ainsi, les mêmes règles s'appliquent en ligne comme hors ligne (Commission européenne 2020b).

La loi NetzDG s'applique aux plateformes dont la fonction principale est de faciliter le partage et la communication publique de contenus avec d'autres d'utilisateurs, ciblant en particulier les plateformes de médias sociaux ayant plus de deux millions d'utilisateurs en Allemagne et excluant les services de messagerie privée. Le choix des plateformes dans le Code de conduite européen est davantage arbitraire, puisque ce sont les plus grands acteurs du numérique, à savoir Facebook, Microsoft, Twitter et YouTube — rejoints par Instagram, TikTok, Jeuxvideo.com, Snapchat et Dailymotion — qui se sont joints de plein gré au « contrat ». Les obligations relatives à la gestion efficace et transparente des plaintes dans la loi NetzDG, de même que des rapports de transparence pour rendre des comptes sur les mécanismes de traitement de contenus, sont le cœur de la législation. La gestion des plaintes impose des délais serrés (24 heures) pour les contenus manifestement illégaux, mais offre une certaine flexibilité – sept jours – pour les contenus dont la légalité peut être complexe à la déterminer.

Alors que la suppression du contenu est faite de façon globale, notamment lorsque les plateformes conviennent que celui-ci viole leurs normes de communautés, le blocage de contenu se fait sur une base géographique, c'est-à-dire qu'il rend le contenu indisponible seulement en Allemagne. Les mécanismes de traitement de signalements sous le Code européen diffèrent quelque peu du modèle allemand. En effet, les signalements reçus par les plateformes signataires sont effectués, d'une part, via des canaux de signalement disponibles pour les utilisateurs généraux et, d'autre part, via des canaux spécifiques disponibles uniquement pour les signaleurs de confiance (*trusted flaggers*). Sous le Code de conduite, le traitement de contenus se fait principalement par les normes et les procédures mises en place par les plateformes, notamment les normes de communauté (*Communuty Guidelines* ou *Communauty Standards*). La modération peut également tenir compte, le cas échéant, des lois nationales transposant la Décision-cadre 2008/913/JAI.

Tableau 5. Typologie des approches de régulation et mesures préconisées ou mises en œuvre concernant les procédures de signalement

Approche de régulation	Obligations	Mesures établies ou mises en vigueur	Conformité par les plateformes	Sanctions
Hard regulation **NetzDG**	Obligation de résultats (définitions d'objectifs et de délais strictes)	Création de 21 nouvelles incriminations pouvant faire l'objet d'un retrait Réforme des procédures de signalement (délais de retrait notamment) Publication de rapports de transparence tous les six mois Amendes en cas de non-conformité	Nouvelles catégories de signalement Embauche et formation de personnels qualifiés Adaptation du dispositif de signalement aux nouvelles exigences Publication en ligne de rapports de transparence	Amendes en cas d'échec à atteindre les résultats ou à respecter les délais
Soft regulation **Code de conduite pour lutter contre la haine en ligne**	Obligation de moyens (définition de mesures à mettre en œuvre)	Mise en place de standards de publication et de procédures de signalement Respect des lois nationales Respect d'un délai maximum de 24 heures 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 (*trus-ted flaggers*)	Aucune

Source : Badouard (2021).

Discussion

Les deux cadres réglementaires analysés dans ce document mettent en œuvre des rationalités de gouvernement des stratégies de régulation qui diffèrent selon leurs degrés de contraintes (Badouard 2021). En effet, le premier cadre est associé à une réglementation stricte (approche *hard regulation*), imposant aux plateformes des obligations de résultats, en plus de prévoir un mécanisme coercitif – notamment des actions en cas de non-conformité aux obligations – pour assurer l'application (*enforcement*) des mesures. Le second, le Code de conduite pour lutter contre les discours de haine illégaux en ligne, peut plutôt être associé à une réglementation plus douce (approche *soft regulation*). Cette dernière privilégie des instruments volontaires et des obligations de moyens (engagements) ne nécessitant pas un contrôle strict de l'application des mesures (Rapp,

Schmid et Wolff 2011; Gorwa 2019a et 2019b; Badouard 2021). Au-delà de ces similarités et de ces différences, nous allons nous concentrer sur les répercussions qu'ils ont eues en matière de lutte contre les contenus préjudiciables.

À la lumière des informations exposées dans le chapitre précédent, il est possible de tirer plusieurs conclusions. Les premières touchent aux hypothèses de recherche. Notre première hypothèse voulant que la loi allemande ait entrainé des suppressions ou blocages excessifs de contenus est infirmée. En effet, le portrait des indicateurs des trois plateformes analysées permet de contraster que le volume de contenus supprimés est faible. À titre d'illustration, la proportion de plaintes ayant entrainé la suppression ou le blocage d'un contenu a atteint le sommet de 35,5 % chez Facebook (tableau 1) dans le second rapport semestriel de 2019. Dans le dernier rapport publié en 2021, ce taux se situe à 26,5 %, ce qui représente seulement 29 % de contenus supprimés ou bloqués. Chez Instagram (tableau 2), la plus forte proportion de plaintes ayant entrainé la suppression ou le blocage d'un contenu a été enregistrée à 46,3 % (représente 47,2 % de contenus) dans le premier rapport de 2020, alors que le dernier rapport indique seulement 12,3 %. Chez YouTube, cette proportion se situe entre 21,8 % (janvier 2019) et 27,1 % (juillet 2018).

La seconde hypothèse voulant que les obligations en vertu d'une approche *soft regulation* (Code de conduite) soient moins respectées est également infirmée. En effet, les plateformes ont mis en place et renforcé leurs règles de communautés pour se conformer à leurs obligations. Elles ont mis en place des mécanismes de signalement également conçus pour les citoyens européens, en plus d'avoir travaillé en étroite collaboration avec les organisations de la société civile. L'on peut toutefois souligner la lacune de la coopération avec les autorités policières compétentes en matière de signalement de discours haineux illégaux.

Notre troisième hypothèse indiquant que les mesures de régulation ont favorisé davantage de transparence sur les pratiques de modération de contenus peut être confirmée, mais avec nuance. Dans le cadre allemand, la publication des rapports de transparence est une obligation législative. L'appréciation des rapports publiés par les plateformes permet d'ailleurs de constater que ceux-ci sont relativement bien détaillés. Cela n'a toutefois pas empêché l'Allemagne d'infliger une amende de deux millions d'euros, en 2019, à Facebook, accusée alors de ne pas avoir communiqué aux autorités allemandes tous les contenus que la plateforme avait supprimés (Untersinger 2019). La pratique des rapports de transparence s'est également répandue ces dernières années, telle que les rapports sur l'application de Standards de la communauté de Facebook ou des règles de communautés de Google. Néanmoins, ces rapports de transparence ne révèlent pas forcément tout; dès lors, la difficulté de l'accès aux données empêche les autorités gouvernementales de certifier les informations fournies par les plateformes (Badouard 2021 : 117).

La quatrième hypothèse stipulant qu'il y aurait une diminution importante du volume de contenus préjudiciables signalés sur les plateformes est également infirmée. En effet, l'évaluation de la mise en œuvre des engagements des plateformes signataires permet de constater que le nombre de contenus signalés dans l'espace européen est passé de 600 en 2016 à 4 364 en 2020, alors que la proportion de contenus supprimés a suivi cette évolution, passant de 28,2 % en 2016 à 59,1 % en 2017, et se stagnant depuis 2018 entre 70 % et 71 %. Le nombre de plaintes reçues, de même que le volume d'infractions signalées en

vertu de la loi NetzDG, n'ont pas cessé d'augmenter en flèche chez les trois plateformes étudiées. On remarque toutefois une baisse du volume de signalements chez Facebook et chez YouTube, mais celle-ci n'est pas significative puisqu'elle est survenue seulement entre juillet 2020 et janvier 2021.

Le chapitre précédent permet ainsi de constater que la régulation des plateformes est possible, qu'elle soit au niveau étatique ou supranational. L'expérience allemande démontre, d'une part, que les États peuvent imposer des règles applicables aux géants numériques à l'intérieur de leurs frontières et, d'autre part, que les plateformes peuvent s'y conformer en ajustant leurs façons de faire. Plusieurs éléments peuvent être soulignés pour témoigner de cette conformité : la publication de rapports de transparence spécialement conçus pour les autorités allemandes et accessibles au public, la mise en place de mécanismes spéciaux de traitement des signalements, l'obligation de supprimer les contenus dans un délai contraignant, la nomination d'un représentant de l'entreprise en Allemagne, etc. La loi NetzDG est en ce sens considérée comme étant « révolutionnaire » (Alkiviadou 2019 : 22):

> because it adopts such a stringent approach to the regulation of hate speech on social media, as demonstrated through the obligation on networks to appoint agents solely for the purpose of this law, as well as the huge potential fines associated with the non-conformity to the legislation by the legislation.

Plusieurs motivations peuvent expliquer la conformité des plateformes à la régulation. Il y a entre autres des intérêts économiques, puisque l'impératif d'affaire de ces acteurs est de fournir du contenu qui soit utile aux utilisateurs. En ce sens, le manquement à la suppression ou au blocage de contenus problématiques peut avoir un effet de perception négative chez le public (Helberger, Pierson et Poell 2018; Badouard 2021). L'engagement des utilisateurs, « créateurs » de contenus, est donc fondamental pour les plateformes.

Par ailleurs, les données du Code de conduite de la Commission européenne sont également encourageantes lorsqu'on analyse leur évolution dans le temps. Il est toutefois possible de dresser quelques limites. En effet, un des principaux défis – voire obstacles – en termes d'application du Code, est que les plateformes dépendent en grande partie des utilisateurs pour signaler les discours de haine. Ainsi, le fonctionnement même du Code dépend grandement des utilisateurs, censés identifier et signaler les discours de haine (Alkiviadou 2019). En outre, la définition du discours de haine utilisée dans la décision-cadre est jugée très large, laissant aux États membres un pouvoir discrétionnaire lors de la transposition des dispositions (Quintel et Ullrich 2020).

> This has led to a different threshold concerning the content that is to be criminalised in the Member States, which complicates any harmonised application of the Framework Decision or the Code of Conduct. Legal provisions prohibiting hate speech may be interpreted loosely and applied selectively, so that there are diverging legal requirements for online platforms in the different Member States (Quintel et Ullrich 2020 : 5).

Cela complexifie donc, pour les plateformes de médias sociaux, l'analyse et la prise en compte des lois nationales dans la détermination de l'illégalité d'un contenu. Or, dans les démocraties occidentales, ce rôle revient souvent aux tribunaux. Quintel et Ullrich (2020) sont d'avis qu'il y a une délégation des tâches des tribunaux nationaux aux acteurs privés. D'autant plus que le Code de conduite ne prévoit pas de mécanisme d'appel pour les contenus supprimés même s'ils n'ont pas été trouvés illégaux. Cela laisse peu d'options aux personnes dont le contenu est présumé illégal d'engager un recours après que leur contenu ait été bloqué ou supprimé.

La lutte contre les discours haineux en ligne représente un défi de taille, tant au niveau juridique qu'au niveau technique. En effet, une des différences marquantes entre la modération de contenus préjudiciables sur les médias traditionnels et celle sur les médias sociaux repose sur le fait que les systèmes de modération employés par les premiers – les médias traditionnels – permettent de « censurer » les propos haineux avant leur publication ou leur diffusion; tandis que les systèmes des médias sociaux reposent surtout sur des utilisateurs déjà offensés (*already-offended users*) (Brown 2018; Ullmann et Tomalin 2020). Comme nous l'avons expliqué, cette logique de modération dite réactive (Bakalis et Hornle 2021; Gillespie 2010 et 2017) comporte des limites : le délai entre la publication du contenu offensant ou répréhensible et sa suppression peut être long; les plateformes ne peuvent donc pas garantir qu'aucun contenu de ce type n'apparaitra ou que les utilisateurs n'y seront pas exposés. La modération proactive, qui implique l'examen du contenu avant sa publication, pose également des défis techniques et humains. D'abord, il y a un souci d'efficacité et d'erreur d'évaluations, puisque les algorithmes ne peuvent pas tout détecter. À titre d'illustration, Google (2021) indique cette complexité dans son rapport de transparence.

> Même si la technologie est devenue très utile pour identifier certains types de contenus controversés (par exemple, pour trouver des objets et des modèles rapidement et à grande échelle dans les images, les vidéos et les contenus audio), l'examen manuel de la situation en contexte reste la méthode plus efficace. Par exemple, les algorithmes ne savent pas toujours faire la différence entre la propagande terroriste, les vidéos sur les droits de l'homme, les discours d'incitation à la haine et les comédies provocatrices. Une intervention humaine est ainsi souvent nécessaire pour prendre la décision finale (Google 2021).

Ensuite, il faudrait des centaines de milliers de modérateurs pour filtrer chaque contenu avant sa publication. Étant donné le nombre de contenus publiés chaque minute dans le monde, et dans plusieurs langues, les délais de publication de contenu seraient étirés. Bref, tous ces défis nous révèlent l'équilibre délicat entre le devoir de garantir la garantir la liberté d'expression et l'impératif de lutter contre les discours préjudiciables en ligne.

Conclusion

La recrudescence de la réglementation des plateformes à l'échelle planétaire témoigne sans doute de la prise de conscience des États de la faiblesse du modèle du « laisser-faire », qui a longtemps caractérisé les règles encadrant ces géants numériques.

Cette note de recherche a tenté modestement d'analyser les répercussions des approches de régulation déjà en vigueur dans d'autres pays. L'analyse comparative de deux modèles de régulation, la réglementation douce ou *soft regulation* et la réglementation stricte ou *hard regulationi,* permet de constater qu'il est possible de poursuivre le même objectif tout en empruntant des moyens différents. Alors que la première est critiquée pour sa « privatisation » de la régulation des contenus aux entreprises privées ou pour son modèle fragilisé dépendant de la bonne volonté des signataires, la seconde est jugée très stricte, faisant craindre la violation de la liberté d'expression.

Notre travail montre, cependant, que ces deux approches portent leurs fruits, et qu'il semble y avoir une certaine collaboration des plateformes concernées à se conformer aux obligations. Évidemment, une telle conformité est sans doute motivée par des incitations variées, mais surtout économiques. Avec la pression grandissante pour une réglementation plus ferme, ces entreprises savent s'adapter, voire devancer les législations. C'est en ce sens que les contenus ou comportements prohibés sur leurs plateformes dupliquent beaucoup ceux qui sont déjà répréhensibles dans beaucoup de pays.

Enfin, l'argument de la liberté d'expression est souvent dressé en épouvantail pour justifier la crainte d'une régulation des contenus préjudiciables. Nous croyons que la liberté d'expression, considérée comme un droit absolu, revient à considérer qu'elle implique également la liberté d'opprimer, d'attaquer et d'offenser les autres. Or, dans le contexte des discours haineux, cette « liberté d'expression » empêche justement les personnes touchées – en l'occurrence des victimes – de s'exprimer. Dès lors, l'équilibre entre la libre expression des uns et la dignité et le droit à la protection des autres peut tout à fait être établi.

Références bibliographiques

Abedi, Maham, « "Tip of the iceberg": Why Canada's online hate-crime data doesn't tell the full story », *Global Newsi*, 2 mai 2019, <https://globalnews.ca/news/5227087/cyber-hate-crime-data-canada/>.

Alexandre, Marie, « La rigueur scientifique du dispositif méthodologique d'une étude de cas multiple », *Recherches qualitatives,* vol. 32, nº 1, 2013, p. 26-56.

Allen, Mary, « Statistiques sur les crimes déclarés par la police au Canada, 2017 », *Statistique Canada*, Ottawa, 2018, 53 p.

Armstrong. Amelia, « Les crimes haineux déclarés par la police au Canada, 2017 », *Statistique Canada*, Ottawa, 2019, 32 p.

Badouard, Romain, « La régulation des contenus sur Internet à l'heure des « fake news » et des discours de haine », *Communications,* vol. 1, nº 1, 2020, p. 161-173, <https://doi.org/10.3917/ commu.106.0161>.

Badouard, Romain, « Modérer la parole sur les réseaux sociaux: Politiques des plateformes et régulation des contenus », *Réseaux,* vol. 1, nº 1, 2021, p. 87-120, <https://doi.org/10.3917 /res.225.0087>.

Bakalis, Chara et Julia Hornle, «The Role of Social Media Companies in the Regulation of Online Hate Speech», *Studies in Law, Politicss, and Society,* vol. 85, 2021, p. 75-100.

Baldauf, Johannes, Julia Ebner et Jakob Guhl, « Hate Speech and radicalisation Online: The OCCI research report », *Institute for Strategic Dialogue*, Londres, 2019, 65 p.

Banks, James, « Regulating hate speech online », *International Review of Law, Computers and Technology,* vol. 24, nº 3, 2010, p. 233-239.

Brown, Alexander, « What is so special about online (as compared to offline) hate speech? », *Ethnicities,* vol. 18, nº 3, 2018, p. 297-326.

Castonguay, Alec, « Chers géants du Web, la récréation est terminée », *L'actualité*, 7 avril 2021, <https://lactualite.com/lactualite-affaires/chers-geants-du-web-la-recreation-est-terminee/>.

CBC News, « Canadians appear to be more hateful online. Here's what you can do about it », 20 janvier 2017, <https://www.cbc.ca/news/canada/marketplace-racism-online-tips-1.3943351>.

Charest, Nicolas, « Veille », dans Côté, Louis et Jean-François Savard (dir.), *Le Dictionnaire encyclopédique de l'administration* publique, 2012, <www.dictionnaire.enap.ca>.

Chetty, Naganna et Sreejith Alathur, « Hate speech review in the context of online social networks », *Aggression and violent behavior,* vol. 40, 2018, p. 108-118.

Claussen, Victor, « Fighting hate speech and fake news. The Network Enforcement Act (NetzDG) in Germany in the context of European legislation », *Rivista Di Diritto Dei Media,* vol. 3, 2018, p. 1-27.

Commission européenne, *Le code de conduite visant à combattre les discours de haine illégaux en ligne,* 2020, 4 p.

Commission européenne, *Le code de conduite visant à combattre les discours de haine illégaux en ligne,* 2020, <https://ec.europa.eu/commission/presscorner/detail/fr/qanda_20_1135>.

Conseil de l'Union européenne, *Décision-cadre sur la lutte contre certaines formes et manifestations de racisme et de xénophobie au moyen du droit pénal (2008/913/JAI)*, 2008, <https://eur-lex.europa.eu/legal-content/fr/ALL/?uri=CELEX%3A32008F0913>.

Cordier, Quentin, « L'économie de plateforme: description d'un phénomène d'intermédiation », *Enjeux et défis juridiques de l'économie de plateforme*, 2019, p. 7-33.

Crawford, Kate et Tarleton Gillespie. « What is a flag for? Social media reporting tools and the vocabulary of complaint », *New Media & Society,* vol. 18, nº 3, 2016, p. 410-428.

Delker, Janosch, « Germany's balancing act: Fighting online hate while protecting free speech », *Politico*, 1er octobre 2020, <https://www.politico.eu/article/germany-hate-speech-internet-netzdg-controversial-legislation/>.

DeNardis, Laura et Andrea M. Hackl, « Internet governance by social media platforms », *Telecommunications Policy*, vol. 39, nº 9, 2015, p. 761-770.

European Commission, *Tackling Illegal Content Online: Towards an enhanced responsibility of online platforms (Communication)*, 2017, 20 p.

European Commission, *Commission Recommendation of 1.3.2018 on measures to effectively tackle illegal content online*, 2018, 17 p.

Facebook, *NetzDG Transparency Report*, juillet 2018, 8 p.

Facebook, *NetzDG Transparency Report*, janvier 2019, 12 p.

Facebook, *NetzDG Transparency Report*, juillet 2019, 13 p.

Facebook, *NetzDG Transparency Report*, janvier 2020, 14 p.

Facebook, *NetzDG Transparency Report*, juillet 2020, 17 p.

Facebook, *NetzDG Transparency Report*, janvier 2021, 17 p.

Flew, Terry, Fiona Martin et Nicolas Suzor, « Internet regulation as media policy: Rethinking the question of digital communication platform governance », *Journal of Digital Media & Policy*, vol. 10, nº 1, 2019, p. 33-50.

Fortin, Marie-Fabienne et Johanne Gagnon, *Fondements et étapes du processus de recherche: méthodes quantitatives et qualitatives*. Chenelière éducation, Montréal, Québec, 2016.

Gasser, Urs et Wolfgang Schulz, « Governance of online intermediaries: Observations from a series of national case studies », *Berkman Center Research Publication*, vol. 18, 2015, 79 p.

Gillespie, Tarleton, « The politics of "platforms" », *New media & society*, vol. 2, nº 3, 2010, p.347-364.

Gillespie, Tarleton, *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*, New Haven, Yale University Press, 2018a.

Gillespie, Tarleton, « Platforms are not intermediaries », *Georgetown Law Technology Review,* vol. 2, nº 2, 2018b, p. 98-216.

Gillespie, Tarleton, « Regulation of and by Platforms », *The SAGE Handbook of Social Media*, 2017.

Guillemin, Gabrielle, « Internet companies alone can't prevent online harms », *Article 19*, 17 avril 2020, <https://www.article19.org/resources/internet-companies-alone-cant-prevent-online-harms/>.

Guerin, Cécile, Zoé Fourel et Cooper Gatewood, « La pandémie de COVID-19 : terreau fertile pour la haine en ligne », *Institute for Strategic Dialogue*, 2021, 24 p.

Gorwa, Robert, « The platform governance triangle: Conceptualising the informal regulation of online content », *Internet Policy Review*, vol. 8, n° 2, 2019a, p. 1-22.

Gorwa, Robert, « What is platform governance? », *Information, Communication & Society,* vol. 22, nº 6, 2019b, p. 854-871.

Gouvernement du Canada, *Initiative conjointe pour la recherche en matière de citoyenneté numérique*, 2020, <https://www.canada.ca/fr/patrimoine-canadien/services/desinformation-en-ligne/initiative-conjointe-recherche-citoyennete-numerique.html>.

Gouvernement du Canada, *Réglementation des plateformes de médias sociaux*, 2020, <https://rechercher.ouvert.canada.ca/fr/qp/id/pch,PCH-2020-QP-00084>.

Gozlan, Angélique, « Facebook: de la communauté virtuelle à la haine », *Topique*, nº 1, 2013, p. 121-134.

Helberger, Natali, Jo Pierson et Thomas Poell, « Governing online platforms: From contested to cooperative responsibility », *The information society*, vol. 34, nº 1, 2018, p. 1-14.

Heldt, Amélie Pia, « Reading between the lines and the numbers: an analysis of the first NetzDG reports », *Internet Policy Review,* vol. 8, nº 2, 2019.

Heldt, Amélie Pia, « Germany is amending its online speech act NetzDG... but not only that », *Internet Policy Review,* 2020, <https://policyreview.info/articles/news/germany-amending-its-online-speech-act-netzdg-not-only/1464>.

Housefather, Anthony, « Agir pour mettre fin à la haine en ligne », *Comité permanent de la justice et des droits de la personne*, 2019, 80 p.

Human Rights Watch, *Germany: Flawed Social Media Law*. 2018, <https://www.hrw.org/news /2018/02/14/germany-flawed-social-media-law>.

Instagram, *NetzDG Transparency Report*, juillet 2019, 12 p.

Instagram, *NetzDG Transparency Report*, janvier 2020, 13 p.

Instagram, *NetzDG Transparency Report*, juillet 2020, 13 p.

Instagram, *NetzDG Transparency Report*, janvier 2021, 13 p.

Johnson, Brett Gregory, « Speech, harm, and the duties of digital intermediaries: Conceptualizing platform ethics », *Journal of Media Ethics*, vol. 32, nº 1, 2017, p. 16-27.

Jørgensen, Rikke Frank et Lumi Zuleta, « Private Governance of Freedom of Expression on Social Media Platforms. *Nordicom Review*, vol. 41, nº 1, 2020, p. 51-67.

Kasakowskij, Thomas, *et al.*, « Network enforcement as denunciation endorsement? A critical study on legal enforcement in social media », *Telematics and Informatics,* vol. 46, 2020, p.1-13.

Keipi, Teo, *et al*., *Online hate and harmful content: Cross-national perspectives*, Londres, Taylor & Francis, 2016.

Knoepfel, Peter, *et al*., *Analyse et pilotage des politiques publiques*, Québec, Presses de l'Université du Québec, 2015.

Krishnamurthy, Vivek, « Planned social media regulations set a dangerous precedent », *The Conversation*, 15 mars 2021, <https://theconversation.com/planned-social-media-regulations-set-a-dangerous-precedent-155844>.

Loi sur le ministère du Patrimoine canadien, L.C., 1995, ch. 11.

Maggetti, Martino, « Hard and soft governance », dans Lynggaard Kennet *et al.* (dir.), *Research methods in European Union studies*, Londres, Palgrave Macmillan, 2015, p. 252-265.

Martens, Berlin, « An economic policy perspective on online platforms », *Institute for Prospective Technological Studies Digital Economy Working Paper*, 2016, 61 p.

Martinez, Paloma, « Les Canadiens veulent un contrôle strict des médias sociaux pour prévenir la haine et le racisme en ligne », *Radio-Canada International*, 25 janvier 2021, <https://www.rcinet.ca/fr/2021/01/25/les-canadiens-veulent-un-controle-strict-des-medias-sociaux-pour-prevenir-la-haine-et-le-racisme-en-ligne/>.

Ministry of Justice and Consumer Protection, *Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act, NetzDG) - Basic Information (2017)*, 2018, <https://www.bmjv.de/DE/Themen/FokusThemen/NetzDG/NetzDG_EN_node.html>.

Noël, Christian, « Un nouveau chien de garde numérique contre la haine en ligne », *Radio-Canada,* 21 février 2021, <https://ici.radio-canada.ca/nouvelle/1769922/chien-garde-numerique-haine-contenus-haineux-internet>.

Obar, Jonathan A. et Steven S. Wildman, « Social media definition and the governance challenge: An introduction to the special issue », *Telecommunications policy*, vol. 39, nº 9, 2015, p. 745-750.

OCDE, « The Role of Internet Intermediaries in Advancing Public Policy Objectives », *OCDE Publishing*, Genève, 2021, 200 p.

Patrimoine canadien, *Raison d'être, mandat et rôle*, 2020, <https://www.canada.ca/fr/patrimoine-canadien/organisation/mandat.html>.

Payette, Julie, *Un Canada plus fort et résilient. Discours du Trône*, 2020, 34 p.

Perset, Karine, « The economic and social role of internet intermediaries », *OCDE*, 2010, 49 p.

Piquard, Alexandre, « La suspension des comptes de Donald Trump et le débat sur la régulation des réseaux sociaux », *Le Monde*, 11 janvier 2021, <https://www.lemonde.fr/economie/article/2021/01/11/comment-reguler-les-reseaux-sociaux-le-cas-trump-relance-le-debat_6065896_3234.html>.

Premier ministre du Canada, *Lettre de mandat du ministre du Patrimoine canadien*, 2019, <https://pm.gc.ca/fr/lettres-de-mandat/2019/12/13/lettre-de-mandat-du-ministre-du-patrimoine-canadien>.

Premier ministre du Canada, *Lettre de mandat supplémetaire du ministre du Patrimoine canadien, 2021*, <https://pm.gc.ca/fr/lettres-de-mandat/2021/01/15/lettre-de-mandat-supplemen taire-du-ministre-du-patrimoine-canadien>.

Quintel, Teresa et Carsten Ullrich, « Self-regulation of fundamental rights? The EU Code of Conduct on Hate Speech, related initiatives and beyond », dans Bilyana Petkova et Tuomas Ojanen (*dir.*), *Fundamental Rights Protection Online*, Cheltenham, Edward Elgar Publishing, 2020.

Rapp, Marc Steffen, Thomas Schmid et Michael Wolff, « Hard or Soft Regulation of Corporate Governance? », *HHL Research Paper Series in Corporate Governance,* nº 6, 2011.

Robillard, Jean-Philippe, « Depuis l'arrivée de Trump, la haine explose sur le Web au Canada », *Radio-Canada*, 2 février 2017, <https://ici.radio-canada.ca/nouvelle/1015082/intolerance-racisme-web-internet-canada-impact-trump-etude-cbc-marketplace>.

Schulz, Wolfgang, « Regulating intermediaries to protect privacy online–the case of the German NetzDG », *Personality and Data Protection Rights on the Internet*, 2018.

Siapera, Eugenia et Paloma Viejo-Otero, « Governing Hate: Facebook and Digital Racism », *Television & New Media,* vol. 22, nº 2*,* 2021, p. 112-130.

Silver, Janet E., « Regulation of online hate speech coming soon, says minister », *iPolitics*. 21 janvier 2021, <https://ipolitics.ca/2021/01/29/regulation-of-online-hate-speech-coming-soon-says-minister/>.

Soilleux-Mills, Anna, Cathryn Hopkins et Butt, Hamzah, « Online harms regulation gains momentum and shape », *Lexology*, 2021, <https://www.lexology.com/library/detail.aspx?g=4c0be777-d91f-4ed5-a48a-d841d6e688bc>.

Statistique Canada, *Données sur les crimes haineux déclarés par la police, 2016*, Ottawa, 2017, 8 p.

Statistique Canada, *Données sur les crimes haineux déclarés par la police, 2017*, Ottawa, 2018, 13 p.

Suzor, Nicolas, « Digital constitutionalism: Using the rule of law to evaluate the legitimacy of governance by platforms », *Social Media+ Society,* vol. 4, nº 3*,* 2018, p. 1-11.

Taylow Wessing, *Online harms: the regulation of internet content*. 2019, <https://www.taylorwessing.com/download/article-online-harms.html>.

Thelle, Martin H. *et al.,* « Online Intermediaries: Impact on the EU economy », *Copenhagen Economics*, 2015, 49 p.

Theil, Stefan, « The Online Harms White Paper: comparing the UK and German approaches to regulation », *Journal of Media Law,* vol. 11, nº 1, 2019, p. 41-51.

Ullmann, Stefanie et Marcus Tomalin, « Quarantining online hate speech: technical and ethical perspectives », *Ethics and Information Technology,* vol. 22, nº 1, 2020, p. 69-80.

Untersinger, Martin, « L'Allemagne inflige à Facebook une amende de 2 millions d'euros en vertu de sa loi sur les réseaux sociaux », *Le Monde*, 2 juillet 2019, <https://www.lemonde.fr/pixels/article/2019/07/02/l-allemagne-inflige-a-facebook-une-amende-de-2-millions-d-euros-en-vertu-de-sa-loi-sur-les-reseaux-sociaux_5484451_4408996.html>.

Wagner, Ben, *et al.*, « Regulating Transparency? Facebook, Twitter and the German Network Enforcement Act », *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, p. 261-271.

Yale, Janet, « L'avenir des télécommunications : le temps d'agir », *Groupe d'examen du cadre législatif en matière de radiodiffusion et de télécommunications*, 2020, 260 p.

YouTube, *Removals under the Network Enforcement Law*, juillet 2018; janvier 2019; juillet 2019; janvier 2020; juillet 2020; janvier 2021, <https://transparencyreport.google.com/netzdg/youtube?hl=en>.

Zipursky, Rebecca, « Nuts About NETZ: The Network Enforcement Act and Freedom of Expression », *Fordham International Law Journal,* vol. 42, nº 4, 2019, p. 1325-1374.

Centre d'études
sur l'intégration
et la mondialisation

Adresse civique :
UQAM, 400, rue Sainte-Catherine Est
Pavillon Hubert-Aquin, bureau A-1560
Montréal (Québec) H2L 2C5 CANADA

Adresse postale :
Université du Québec à Montréal
Case postale 8888, succ. Centre-Ville
Montréal (Québec) H3C 3P8 CANADA

Téléphone : 514 987-3000, poste 3910
Télécopieur : 514 987-0397
Courriel : ceim@uqam.ca
Site web : www.ceim.uqam.ca

Auteur

Dorian Mouketou est diplômé de la maitrise en administration publique de l'École nationale d'administration publique (ENAP), Montréal.

dorianpaterne.mouketou@enap.ca