Université de Montréal

Modération de contenu sur Internet

Analyse de l'approche proposée en 2021 par le gouvernement du Canada pour lutter contre le contenu préjudiciable

Note : avec le consentement de l'autrice, le département de communication donne accès à ce document à titre d'exemple d'un format de travail dirigé. Ce travail dirigé consiste en une analyse de politique publique du gouvernement du Canada. Sa spécificité tient au fait qu'il se termine par une « simulation d'un mémoire » qui aurait pu être présenté dans le cadre d'une consultation publique.

par
Véronique Bolduc

Unité académique du département des communications
Faculté des arts et des sciences

Travail dirigé présenté en vue de l'obtention du grade
de maîtrise en sciences de la communication,
option communication politique

Juillet 2022

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE D'ABRÉVIATIONS

NetzDG : Network Enforcement Act

SCL : Service de communication en ligne

~~~~~~~~~~~~~~~~~~~~

Pour des fins de lisibilité et de concision, je référerai aux documents suivants, composant mon corpus de textes, en les nommant directement, plutôt qu'en citant leurs références :

- Approche (L'approche proposée par le gouvernement du Canada pour lutter contre les contenus préjudiciables en ligne)
- *Guide de discussion* (Ministère du Patrimoine canadien, 2021a);
- *Document technique* (Ministère du Patrimoine canadien, 2021b);
- *Rapport synthèse* (Ministère du Patrimoine canadien, 2022a).
~~~~~~~~~~~~~~~~~~~~

1 INTRODUCTION

Dans les dernières années, les environnements numériques semblent être devenus de plus en plus hostiles étant donné la visibilité accrue de contenus préjudiciables en ligne. Que ce soit la diffusion en direct de la tuerie de Christchurch en Nouvelle-Zélande le 15 mars 2019, la propagande terroriste de l'État islamique pour recruter des membres ou encore les fausses nouvelles qui incitent à la violence comme le cas de l'assaut du Capitole aux États-Unis le 6 janvier 2021, nombreux exemples de contenus préjudiciables circulent en ligne et semblent rendre Internet un lieu moins sécuritaire et moins inclusif. À travers les années, une pression grandissante est mise sur les gouvernements pour qu'ils agissent en matière de modération de contenu puisque les plateformes ne semblent pas y arriver seules. Toutefois, les questions de modération de contenu soulèvent des problématiques vastes et complexes. Ces décisions, prises en quelques secondes, soit par des humains, par des algorithmes ou les deux conjointement, ont des impacts importants au niveau démocratique, culturel, légal et plusieurs autres.

Durant l'été 2021, le gouvernement du Canada a lancé une consultation publique dans le but de proposer éventuellement un nouveau cadre réglementaire et législatif pour lutter contre le contenu préjudiciable en ligne. Cette proposition avait pour but de réglementer les « services de communication en ligne » (SCL) (*Guide de discussion*), incluant, entre autres, les médias sociaux. Ce cadre aurait ciblé les cinq catégories de contenus préjudiciables les plus graves soit le contenu terroriste, le contenu incitant à la violence, les discours haineux, le partage non consensuel d'images intimes et le contenu d'exploitation sexuelle d'enfants en ligne (Ministère du Patrimoine canadien, 2021c, §1). L'objectif est de rendre Internet sûr et sécuritaire en plus de tenter de remédier au caractère arbitraire qui semble guider les pratiques de modération de contenu actuelles en réclamant des rapports de transparence.

Mon travail dirigé consistera à faire une analyse de cette Approche[1]. D'abord, en guise de problématique, je développerai un portrait de la modération de contenu, présenterai la situation du Canada concernant le contenu préjudiciable et ferai un bref survol de cadres réglementaires similaires à l'international. Ensuite, en me concentrant sur les documents fournis lors de la

[1] Pour des fins de lisibilité et de concision, le terme Approche sera utilisé pour désigner l'approche proposée par le gouvernement du Canada pour lutter contre les contenus préjudiciables en ligne.

consultation publique, soit le *Guide de discussion*[2] et le *Document technique*[3], ainsi que le *Rapport synthèse*[4] de la consultation, j'analyserai quatre axes principaux de l'Approche : les entités réglementées, les catégories de contenus préjudiciables réglementées, les nouvelles règles et obligations et les nouveaux organismes réglementation. Pour finir, je présenterai un sommaire des résultats de l'analyse sous la forme d'une simulation d'un mémoire qui pourrait être soumis au gouvernement pour une consultation publique.

1.1 Contexte : l'approche proposée par le gouvernement du Canada

Le 29 juillet 2021, le ministère de Patrimoine canadien, qui s'occupe d'affaires concernant la vie culturelle, civique et économique des Canadiens, a proposé une approche pour lutter contre le contenu préjudiciable en ligne, désigné ci-après sous le terme simple de l'Approche. L'Approche a été proposée à la suite d'une pression grandissante des Canadiens et Canadiennes de voir des actions concrètes face aux contenus haineux et illégaux qui circulent en ligne (Ministère du Patrimoine canadien, 2021c). Cette approche aurait visé l'instauration d'un cadre législatif et réglementaire[5] pour obliger les SCL à être plus sévères, constants et transparents dans leurs décisions de modération de contenu afin de lutter contre les contenus préjudiciables. L'Approche a été présentée sous la forme de deux documents en format page Web : Le *Guide de discussion* et le *Document technique*. Le *Guide de discussion* présente un portrait général des motivations, objectifs et changements légaux que l'Approche aurait apportés. Le *Document technique* explique plus en détail la mise en application et le fonctionnement concret de l'Approche. Du 29 juillet au 25 septembre 2021, l'Approche a été soumise à une consultation publique afin de récolter les observations et remarques de la population canadienne. À travers ce processus de consultation, le gouvernement du Canada a reçu 422 réponses au total, provenant de particuliers (350 réponses), d'organisations de la société civile (39 réponses), de l'industrie (19 réponses), du milieu universitaire (13 réponses) et d'organisations gouvernementales (2 réponses). La longueur des réponses varie grandement pouvant être aussi courte que 5 pages ou être aussi longue qu'une

[2] Pour des fins de lisibilité, je réfère au document suivant en le nommant directement, plutôt qu'en le citant : *Guide de discussion* (Ministère du Patrimoine canadien, 2021a).
[3] Pour des fins de lisibilité, je réfère au document suivant en le nommant directement, plutôt qu'en le citant : *Document technique* (Ministère du Patrimoine canadien, 2021b).
[4] Pour des fins de lisibilité, je réfère au document suivant en le nommant directement, plutôt qu'en le citant : *Rapport synthèse* (Ministère du Patrimoine canadien, 2022a).
[5] Un cadre législatif et réglementaire représente l'ensemble des lois, règles, obligations et exigences mises en place pour réguler et encadrer les comportements ou les activités d'entités ou d'individus (Rabeau, s.d., p. 5).

cinquantaine de pages. Le 3 février 2022, un rapport synthèse de la consultation publique a été publié (*Rapport synthèse*).

Selon le *Guide de discussion*, l'Approche avait pour but de conduire à un projet de loi qui aurait été présenté à l'automne 2021. Ce projet de loi devait également s'inscrire dans une stratégie plus large contre les discours haineux, au sein de laquelle aurait aussi fait partie le projet de loi C-36[6]. Cependant, le projet de loi C-36 est mort au feuilleton tout comme cette Approche étant donné la dissolution du gouvernement le 15 août 2021, en vue des élections fédérales d'octobre 2021[7].

Après quelques mois de silence sur ce sujet, le gouvernement libéral, dans les 100 premiers jours de son mandat, a recommencé à travailler sur cette approche depuis le mois de février 2022. Le ministère du Patrimoine canadien a ainsi publié un rapport synthèse de la consultation publique qui avait eu lieu durant l'été 2021 et depuis le 30 mars 2022, un comité consultatif d'experts a été mis sur pied pour réviser l'approche proposée par le gouvernement du Canada et prodiguer des conseils pour son amélioration (Ministère du Patrimoine canadien, 2022 b, §3).

1.2 Un bref survol de l'Approche

Afin d'assurer une meilleure compréhension, cette section présentera un bref survol de l'Approche. Les détails seront explicités dans la section analyse de mon travail dirigé. Pour commencer, l'Approche a pour objectif « […] de soutenir un environnement en ligne sûr, inclusif et ouvert » (*Guide de discussion*) Selon le *Guide de discussion*, les médias sociaux et autres services en ligne sont trop souvent utilisés pour la diffusion de messages haineux, la propagande terroriste ou le partage non consensuel d'images intimes. Le gouvernement considère que les SCL ont déjà les

[6] Le projet de loi C-36 avait été déposé à la Chambre des communes le 23 juin 2021. En bref, ce projet de loi, nommé *Loi modifiant le Code criminel, la Loi canadienne sur les droits de la personne et apportant des modifications connexes à une autre loi (propagande haineuse, crimes haineux et discours haineux)*, visait à tenir responsables les services de communications en ligne « […] du contenu haineux publié dans leur écosystème. » (La Presse canadienne, 2021a, §7) Parmi les changements proposés, une nouvelle définition plus étroite de la haine avait été proposée ainsi qu'un système de recours pour les citoyens qui souhaitent émettre des plaintes concernant des discours haineux.

[7] L'Approche ainsi que le projet de loi C-36 sont des initiatives distinctes des projets de loi C-10 et C-11 qui ont été grandement médiatisés en 2021. Bien qu'ils aient tous un objectif de réguler les géants du Web, le projet de loi C-11 (connu sous le nom de Projet de loi C-10 avant l'élection fédérale de 2021) tentait de soumettre les géants de Web aux lois du CRTC, de réguler les algorithmes de découvrabilité et de modifier le mandat du CRTC (La Presse canadienne, 2021b ; Godbout, 2022).

outils et les moyens mis en place pour modérer le contenu. Ces derniers, cependant, ne le font tout simplement pas assez fermement, prennent rarement en compte l'intérêt du public et n'ont aucune obligation de conserver des preuves ou d'aviser les autorités lorsqu'ils sont témoins d'un acte criminel. En conséquence, l'Approche vise à dicter comment SCL doivent traiter et modérer les contenus préjudiciables.

L'Approche est présentée en deux modules principaux dans lesquels se trouvent des sous-sections :

- Module 1 : Un nouveau cadre législatif et réglementaire (*Guide de discussion*)
    a. « Nouveau cadre législatif et réglementaire » (*Document technique*)
    Cette sous-section explique les définitions des catégories de contenus préjudiciables ciblées (le contenu terroriste, le contenu incitant à la violence, les discours haineux, le partage non consensuel d'images intimes et le contenu d'exploitation sexuelle des enfants) et définit les entités réglementées.
    b. « Nouvelles règles et obligations » (*Document technique*)
    Cette sous-section explique les nouvelles exigences légales auxquelles seront soumis les entités réglementées.
    c. « Établissement des nouveaux organismes de réglementation » (*Document technique*)
    Cette sous-section présente les nouveaux organismes de réglementation, leurs compositions, leurs mandats et leurs fonctionnements. Les nouveaux organismes de réglementation sont le Commissaire à la sécurité numérique, le Conseil de recours en matière numérique du Canada et le Comité consultatif d'experts.
    d. « Pouvoirs réglementaires et pouvoirs d'exécution » (*Document technique*)
    Cette section présente les pouvoirs octroyés aux différents organismes de réglementation pour l'application de l'Approche et les sanctions prévues pour les SCL dans des cas de non-conformité.
- Module 2 : Modifier le cadre législatif canadien existant (*Guide de discussion*)
    Cette section explique comment les lois déjà existantes au Canada seront modifiées et impactées avec la mise en place de cette Approche.

2 PROBLÉMATIQUE

Avant d'entamer l'analyse, un survol de la littérature au sujet de la modération de contenu en ligne était de mise puisque l'Approche aurait régulé les processus de modération de contenu des SCL. Je considère qu'il est important de comprendre les mécanismes de modération de contenu employés par les SCL et les médias sociaux pour pouvoir proposer une approche réglementaire fonctionnelle et adaptée. En ayant une meilleure compréhension des processus de modération de contenu, il est possible de reconnaître les problématiques et les impacts concrets que pourrait avoir un cadre réglementaire sur les environnements numériques et leurs usagers.

Cette section présentera également des données concernant les différentes catégories de contenus préjudiciables au Canada et des approches réglementaires alternatives proposées à l'international. Afin de mieux comprendre la raison du gouvernement de s'attaquer aux catégories de contenus préjudiciables ciblées par l'Approche, je considère qu'il est important d'illustrer l'impact et la présence de ces contenus dans les environnements numériques du Canada. Pour ce qui est de survoler les autres approches de régulation en matière de modération de contenus préjudiciables, ce regard à l'international peut indiquer certaines problématiques, permet d'anticiper des risques et de souligner les solutions alternatives envisageables pour un cadre réglementaire au Canada. Enfin, cette section se terminera avec la présentation des objectifs de recherche de mon travail dirigé.

2.1 Les recherches sur la modération de contenu sur les médias sociaux

Pour réaliser le recensement des écrits, j'ai effectué mes recherches principalement en anglais sur Google Scholar et sur la base de données de la bibliothèque de l'Université de Montréal. J'ai favorisé l'anglais, car les sources en français étaient assez limitées. Dans la littérature scientifique, les résultats en français s'éloignent des médias sociaux et se concentrent davantage dans les domaines des politiques publiques (Badouard, 2021 ; Mouketou, 2021), des entreprises, du marketing ou du droit. À travers l'utilisation de mots clés comme « content moderation », « social media » et « regulation », certains auteurs plus prolifiques de ce domaine comme Gillespie (2020), Roberts (2016 ; 2017 ; 2018), Myers West (2018) et Gerrard (2018) revenaient constamment dans les premiers résultats de recherche. Ensuite, considérant que mon travail dirigé consiste à réaliser une analyse de politique publique, je souhaitais trouver quelques analyses de ce genre au sujet de

cadre réglementaire de modération de contenu. En plus d'être assez limité, aucun auteur ne se distinguait et la majorité des articles trouvés provenaient de journaux spécialisés en droit.

À première vue, les articles semblaient tous reprendre les mêmes thèmes, notamment la modération par automation, surtout les biais d'algorithmes (Binns *et al.*, 2017 ; Gerrard, 2018 ; Gillespie, 2020 ; Gorwa *et al.*, 2020) et une focalisation sur les géants du Web comme Facebook, Twitter et YouTube (Crawford et Gillespie, 2016 ; Gerrard, 2018 ; Jhaver *et al.*, 2018). Par exemple, Gillespie (2020) écrit davantage sur la modération de contenu algorithmique et l'intelligence artificielle, alors que Roberts (2016 ; 2017 ; 2018) traite de la modération de contenu commerciale. Myers West (2018), Jhaver *et al.* (2018), Cook *et al.* (2021) et Riedl *et al.* (2021) étudient le point de vue des usagers face à la modération de contenu. Ces études me semblent pertinentes dans le cadre de mon travail dirigé puisqu'il est pertinent de comprendre davantage l'opinion des usagers sur ces questions afin d'élaborer un cadre réglementaire qui prend en considération leurs besoins et leurs attentes tout en ayant une solution efficace et réaliste.

L'étude de Myers West (2018) a, par exemple, conclu que la plupart des usagers ressentent de la frustration due au manque de transparence au niveau de la modération de contenu et, par conséquent, s'imaginent des théories pour expliquer les mécanismes de modération de contenu. Jhaver *et al.* (2018) s'intéressent à l'impact des « blocklists » sur Twitter au sujet du harcèlement en ligne. Une « blocklist » est un outil externe à Twitter qui crée des listes de comptes « dignes d'être bloqués » (Jhaver *et al*., 2018, p. 3). Les usagers peuvent utiliser cet outil pour bloquer automatiquement plusieurs comptes sans avoir à faire le travail manuellement. À la suite d'entrevues avec des personnes qui utilisent le « blocklist » pour se protéger et avec des usagers qui sont sur cette liste, les auteurs proposent de nouvelles façons de répondre aux besoins des usagers en matière de harcèlement en ligne en rendant disponible, par exemple, des ressources pour les victimes de harcèlement ou en ayant des mécanismes d'appel pour les personnes qui se font bloquer (Jhaver *et al*. 2018, p. 26). Cook *et al.* (2021) soulèvent trois conclusions intéressantes, soit 1) que les usagers ne voient pas vraiment de différence entre les modèles de modération de contenu, 2) qu'ils modèrent par eux-mêmes les comportements toxiques lorsque la plateforme les inclut dans le processus et 3) qu'ils apprécient le fait d'avoir une figure d'autorité qui prend les décisions. Enfin, Riedl *et al.* (2021) s'intéressent à la perception du public américain sur la

modération de contenu sur les médias sociaux. L'article se penche sur deux points en particulier : l'engouement pour la modération de contenu par les plateformes et l'engouement pour des cadres réglementaires mis en place par le gouvernement pour la modération de contenu sur les médias sociaux.

Ensuite, l'étude de Caplan (2018), intitulée « Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches », est largement citée dans les articles récents de ce domaine. Cette étude offre un recensement des différents modèles de modération de contenu par les plateformes : artisanal (cas par cas), communautaire (par les usagers) et industriel (utilisation d'outils d'automation). En plus de présenter un portrait détaillé des différents modèles, Caplan (2018) explique que les plateformes de médias sociaux cherchent constamment à trouver un équilibre entre « avoir une sensibilité au contexte » et « être constant » dans les décisions prises. Avoir une sensibilité au contexte signifie de considérer comment le contenu est reçu dans une culture précise avant de prendre des décisions de modération, alors que le fait d'être constant requiert des règles strictes et inflexibles qui s'appliquent à l'ensemble de la plateforme. Son analyse démontre aussi comment la taille, les valeurs et la mission d'une plateforme auront un impact sur ses décisions face au modèle de modération de contenu. Inspiré du travail de Caplan (2018), Renaissance numérique (2020), un *think tank* français, a publié un guide vulgarisant la modération de contenu pour l'ensemble de la population, mais qui s'adresse principalement aux décideurs de politiques publiques en lien avec les SCL. Cela dit, le guide va au-delà de l'étude de Caplan (2018) en faisant des liens avec les politiques publiques de la France. Le but est d'offrir des indicateurs concrets pour tenter de mieux comprendre l'écosystème interconnecté des plateformes et pouvoir prendre de meilleures décisions face aux cadres législatifs et réglementaires proposés.

2.2 Portrait des contenus préjudiciables en ligne au Canada

Les lignes qui suivent tentent de donner un portrait de la situation au Canada par rapport aux cinq catégories de contenus préjudiciables ciblées par l'Approche, soit le partage non consensuel d'images intimes, le contenu d'exploitation sexuelle d'enfants, le contenu terroriste, le contenu incitant à la violence et les discours haineux, afin de mieux comprendre la décision du gouvernement de les avoir sélectionnées. Notons, le manque flagrant de données récentes, ou du

moins l'accessibilité de ces données, a un impact sur le réalisme du portrait que je tente de dresser dans cette section.

2.2.1 Le partage non consensuel d'images intimes

Pour commencer, au niveau du partage non consensuel d'images intimes au Canada, le dernier rapport gouvernemental trouvé qui se penchait sur cette problématique date d'il y a 9 ans. Déjà en 2013, le groupe de travail du comité de coordination des hauts fonctionnaires sur le cybercrime (2013) reconnaissait qu'il existait de grandes lacunes dans le Code criminel concernant le partage non consensuel d'images intimes (p. 2). J'ai aussi trouvé un rapport portant sur les comportements des jeunes Canadiens face au partage d'images intimes en ligne. Ce rapport mentionne qu'un jeune sur six affirmait que ses images intimes envoyées en textos avaient été partagées (Johnson *et al*., 2018, p. 9). Cependant, il n'est pas possible de savoir si ce partage s'est effectué en ligne ou par un moyen physique (montrer l'image à un ou une ami.e). Dans le cas des « sextos » partagés à une tierce personne, le partage public de ces images est le mode de partage le moins fréquent (Johnson *et al*., 2018, p. 10). Il est à noter toutefois qu'il y a trop peu de données pour comprendre l'ampleur véritable de ce problème au niveau du Canada.

2.2.2 L'exploitation sexuelle des enfants en ligne

Au niveau de l'exploitation sexuelle des enfants en ligne, la Sécurité du Canada est responsable de la stratégie nationale à ce sujet depuis 2004 (Sécurité publique Canada, 2022, § 3). De plus, en 2012, le Canada a ajouté au Code criminel le fait qu'il est illégal d'organiser une infraction d'exploitation sexuelle d'enfants à l'aide d'outils de télécommunications (Centre canadien de protection de l'enfance, 2016, p. 7). Cela dit, plusieurs initiatives de collaboration et d'exigences légales existent déjà entre le gouvernement et l'industrie numérique afin de lutter contre le contenu d'exploitation sexuelle des enfants en ligne. L'Approche complémente le tout en imposant un processus plus robuste de signalement de ce contenu et de transmission des données afin de pouvoir arrêter les délinquants (*Guide de discussion*). Malgré tous les efforts déployés pour lutter contre ces contenus préjudiciables, les chiffres ne cessent d'augmenter. En 2015, cyberaide.ca a reçu 37 352 signalements de contenu d'exploitation sexuelle d'enfant en ligne (Centre canadien de protection de l'enfance, 2016, p. 6) et depuis mars 2020, il y aurait eu une augmentation de 88 % des signalements (Gouvernement du Canada, 2021).

Le projet Arachnid a soulevé des faits intéressants concernant la présence et la circulation de ce type de contenu en ligne à l'échelle internationale. Ce projet est un robot automatisé qui fouille le Web à la recherche d'images d'exploitation sexuelle d'enfants. Lorsqu'il en trouve, des signalements sont automatiquement envoyés auprès des fournisseurs de services électroniques afin qu'ils suppriment le contenu (Centre canadien de protection de l'enfance, 2021, p. 2). Dans leur rapport, publié après trois ans d'activités, le projet a constaté que les fournisseurs de services électroniques ne sont pas fiables au niveau du retrait d'images d'exploitation sexuelle d'enfants (Centre canadien de protection de l'enfance, 2021, p. 2). D'abord, le délai médian de suppression de ce genre de contenu est de 24 h, mais le délai peut aller jusqu'à sept (7) semaines pour certains. Selon le Centre canadien de protection de l'enfance (2021), il est nécessaire d'avoir une obligation réglementaire afin de donner une mesure incitative aux fournisseurs de services électroniques de bloquer ou supprimer les contenus dans de meilleurs délais pour éviter les préjudices envers les victimes (p. 4).

Ensuite, environ 48 % des images signalées par le projet avaient déjà été signalées auparavant au même fournisseur de service électronique. En d'autres mots, les images semblent être publiées à nouveau à la suite des suppressions. Ceci laisse penser que les fournisseurs de services électroniques n'utilisent soit pas d'outils d'automation pour préventivement (*ex ante*) bloquer les images ou qu'ils n'ajoutent pas les images signalées à une base de données pour que leurs outils les bloquent (Centre canadien de protection de l'enfance, 2021, p. 3). Enfin, pour le contenu d'exploitation sexuelle des enfants, il est important de noter que ce contenu ne se trouve pas uniquement dans le Web clandestin (*dark web*). La plupart des images sont hébergées sur le Web « visible » et ensuite partagées sur le Web clandestin (Centre canadien de protection de l'enfance, 2021, p. 7).

2.2.3 Le contenu terroriste, le contenu incitant à la violence et les discours haineux

Le contenu terroriste, le contenu incitant à la violence et les discours haineux sont très souvent pris comme un tout commun. Considérant que l'Approche aurait fait partie d'une stratégie plus large pour lutter contre la haine, il est important de s'attarder sur les impacts des discours haineux au Canada. Dû au manque de réglementations et de lois, « […] online expression of hatred, xenophobia, misogyny, racism, and misinformation have been exponentially amplified and

tolerated to an extent that would never be permitted in other arenas. » (Canadien Citizens' Assembly on Democratic Expression, 2021, p. 31) Entre 2010 et 2017, il y aurait eu 374 cas de crimes haineux en ligne traités par la police au Canada (Housefather, 2019, p. 20). Toutefois, les experts s'entendent pour dire que ce chiffre est loin de représenter l'ampleur réelle de la situation. Plusieurs facteurs expliqueraient le manque de signalements concernant la haine (surtout celle en ligne) comme notamment le fait que les victimes craignent parfois le traitement de la police, ne savent pas ce qui est considéré comme un discours ou crime haineux et d'autres ne comprennent simplement pas le processus de signalement (Houefather, 2019, p. 21). Parmi les données disponibles, les groupes les plus touchés sont la population musulmane (17 %), la population juive (14 %), la population noire (10 %) et la communauté LGBTQIA+ (15 %) (Housefather, 2019, p. 23).

Le plus grand défi dans cette lutte contre les discours haineux est la nécessité de trouver un équilibre entre la liberté d'expression et le respect des droits et libertés des victimes (Housefather, 2019, p. 11). Au-delà de ce défi, les experts affirment la nécessité d'avoir un cadre réglementaire qui s'applique à tous les SCL, une définition claire et accessible de la haine en ligne et des mécanismes de signalement facile à utiliser (Housefather, 2019).

En bref, le Canada n'est pas à l'abri de la circulation des contenus préjudiciables en ligne. Cependant, le manque de données ne permet pas de dresser un portrait adéquat de l'ampleur de la situation. Donc, les catégories de contenus préjudiciables ciblées par le gouvernement du Canada semblent, en effet, problématiques au niveau des préjudices causés auprès de la population et nécessitent de s'y attarder.

2.3 Survol d'approches à l'international

Le Canada n'est pas le premier pays à proposer un cadre législatif et réglementaire pour lutter contre les discours haineux et le contenu illégal en ligne. Depuis plusieurs années, différents pays proposent des initiatives ou des projets de loi pour tenter de limiter la propagation de contenu illégal sur Internet. Toutefois, il y a différents types de gouvernance dépendamment du cadre proposé. Badouard (2021) mentionne qu'il existe soit l'avenue de la « soft » ou de la « hard » gouvernance (p. 89).

D'un côté, la « soft » gouvernance signifie que l'État s'entend sur des engagements publics avec les entités concernées sans toutefois qu'il y ait de mise en application stricte comme des sanctions (Badouard, 2021, p. 90). Il s'agit d'une relation basée sur la confiance entre les entités concernées. Le code de conduite visant à combattre les discours de haine illégaux en ligne de l'Union européenne (Commission européenne, 2016) est un exemple de « soft » gouvernance (Badouard, 2021, p. 89). Plusieurs compagnies informatiques ont accepté d'y adhérer comme, entre autres, Facebook, Microsoft, Twitter, YouTube, Instagram, Snapchat, Dailymotion, TikTok et LinkedIn (Commission européenne, 2016, p. 1). Il s'agit d'engagements publics que les compagnies suivent dans le meilleur de leurs capacités. Parmi les engagements, les compagnies s'engagent, notamment, à mettre en place des processus clairs et précis pour notifier et réviser les contenus signalés, d'essayer de retirer les contenus à l'intérieur de 24 h lorsqu'il s'agit de contenus illégaux, de partager entre elles les meilleures pratiques de modération et d'éduquer les usagers sur les types de contenus qui ne peuvent pas être publiés sur leurs plateformes (Commission européenne, 2016, p. 2). Ce code de conduite ne prévoit aucune obligation, exigence ou sanction et les compagnies y adhèrent sur une base volontaire. Il ne s'agit en rien d'un cadre réglementaire obligatoire à suivre.

De l'autre côté, la « hard » gouvernance consiste à ce que l'État mette en place des obligations et exigences envers des entités réglementées. L'État aura une surveillance stricte de la conformité des lois et règlements, exigera des résultats et pourrait prévoir des sanctions dans des cas de non-conformité (Badouard, 2021, p. 89). Parmi des initiatives de « hard » gouvernance, on trouve le Network Enforcement Act (NetzDg) en Allemagne, la loi Avia en France (Badouard, 2021, p. 89) et le Online Safety Act en Australie.

D'abord, le *Network Enforcement Act* (NetzDg) est entré en vigueur le 1er janvier 2018 (Badouard, 2021, p. 91). Cette loi s'applique aux plateformes qui sont utilisées pour le partage public de contenu. Les communications privées et les compagnies ayant moins de 2 millions d'usagers allemands sont exclues de ce cadre réglementaire. Les entités réglementées doivent aussi mettre en place des mécanismes pour gérer les signalements de contenus illégaux. La loi prévoit que des amendes allant jusqu'à 50 millions € puissent être données aux compagnies de médias sociaux ou aux fournisseurs de services de communication qui ne retirent pas les contenus clairement illégaux

dans une période de 24 h (Center for Democracy & Technology, 2017). Pour le contenu qui n'est pas clairement illégal, les entités réglementées ont un délai de sept jours afin de leur donner plus de temps d'analyse (Center for Democracy & Technology, 2017). Cette loi a aussi déclaré « […] 21 nouvelles incriminations pouvant faire l'objet d'un retrait […] » (Badouard, 2021, p. 92) comme entre autres, la diffamation, le partage non consensuel d'images intimes et l'incitation publique au crime (Center for Democracy & Technology, 2017). Enfin, des rapports de transparence sont requis et des évaluations mensuelles des mécanismes de traitement des plaintes sont obligatoires.

Un deuxième exemple de « hard » gouvernance est la loi Avia qui visait à lutter contre le contenu haineux sur Internet présentée en juin 2020 en France (République française, 2020). Cette loi avait des dispositifs similaires à l'approche proposée par le Canada comme entre autres le retrait de contenus illégaux en 24 h et des amendes pouvant aller jusqu'à 6 % du chiffre d'affaires des compagnies qui ne respectent pas les exigences et obligations (Goosz, 2021). De plus, la loi prévoyait que les entités réglementées doivent retirer en une heure le contenu terroriste et pédopornographique (Vie publique, 2020). Toutefois, même si la loi a été adoptée en France, plusieurs dispositifs, tels que les délais de retraits, ont été jugés comme portant atteinte à la liberté d'expression par le Conseil constitutionnel et ont été retirés (Kieffer et Laurent, 2020).

Enfin, le troisième exemple de « hard » gouvernance est le Online Safety Act, adopté en Australie en 2021. L'objectif de la loi est d'améliorer la sécurité en ligne des Australiens et s'applique aux médias sociaux, certains services électroniques et les services Internet (The Parliament of Australia, 2021). Le eSafety Commissioner fut créé et son rôle est de s'occuper, entre autres, des plaintes d'intimidation en ligne envers des enfants, des plaintes concernant les « cyber-abus » visant des adultes et des plaintes de partages non consensuels d'images intimes (The Parliament of Australia, 2021). Les entités réglementées devront aussi fournir des rapports de transparence sur la façon dont elles répondent aux exigences et obligations et le eSafety Commissionner pourrait avoir recours à des pénalités civiles dans des cas de non-conformité (eSafety Commissioner, 2022). Ces trois exemples de régulations sont considérés de la « hard » gouvernance par l'État puisque chaque loi prévoit des sanctions si les exigences et obligations ne sont pas respectées. De plus, ces trois lois demandent que les entités réglementées fournissent des rapports de transparence, des rapports de résultats, sur les moyens utilisés pour modérer le contenu.

Badouard (2021) semble, en effet, placer davantage ces options de gouvernance sur un spectre plutôt qu'une simple dualité puisqu'il considère que le cadre réglementaire du Royaume-Uni se situe entre la « soft » et la « hard » gouvernance. Durant l'hiver 2022, le Royaume-Uni a déposé à la Chambre des communes une nouvelle version d'un projet de loi nommé le Online Safety Bill (Department for Digital, Culture, Media & Sport et Dorries, 2022). Ce projet de loi est présenté comme étant un cadre réglementaire qui assurera des environnements en ligne plus sains et sécuritaires. Concrètement, cette loi tente de limiter l'exposition au contenu illégal en ligne et limiter la visibilité de la pornographie pour les enfants, tout en préservant la liberté d'expression. La loi s'attaque autant au contenu illégal, mais aussi au contenu considéré préjudiciable, mais légal (« harmful but legal ») comme le harcèlement, l'automutilation et les discours encourageant les troubles alimentaires (Department for Digital, Culture, Media & Sport et Dorries, 2022). En plus du retrait de ce contenu, des mécanismes d'appel devront être mis en place pour que les usagers puissent contester les décisions prises par les plateformes. Enfin, la loi prévoit rendre légalement responsables les « hauts placés » des compagnies pour ce qui a rapport avec la transmission et la conservation de données. Selon Badouard (2021), il s'agirait d'une « […] voie intermédiaire, celle d'une obligation de moyens (et non de résultats), assortie d'un contrôle strict, comprenant une menace d'amendes si les plateformes n'obtempèrent pas. » (p. 90) Je présente, dans la figure 1, une synthèse des types de gouvernance basée sur le texte de Badouard (2021).

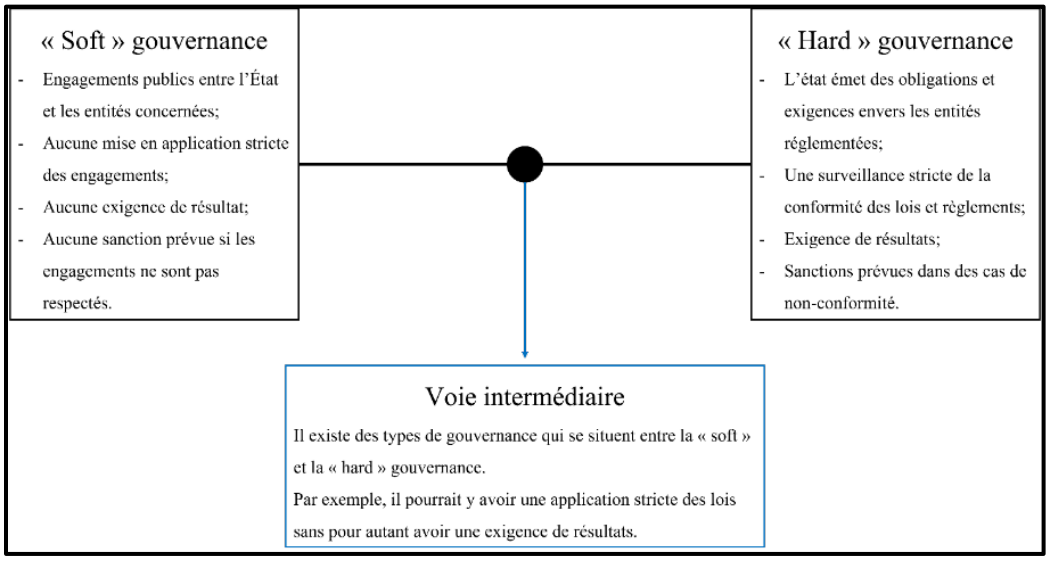

Figure 1 Le spectre des types de gouvernance (Basé sur le texte de Badouard [2021])

L'Approche semble s'inscrire dans une initiative de « hard » gouvernance. Le gouvernement met en place des organismes de réglementation pour assurer une application stricte des nouvelles règles et obligations (*Guide de discussion*). Ensuite, le gouvernement exige aussi des résultats comme l'obligation de rendre le contenu inaccessible au Canada à l'intérieur d'un délai prédéterminé. Finalement, des sanctions assez strictes, allant d'amendes monétaires au blocage de certains sites (*Guide de discussion*) s'ajoutent aux caractéristiques d'un cadre réglementaire de « hard » gouvernance. Je reviendrai plus en détail sur les différents éléments de l'Approche dans le chapitre d'analyse de mon travail dirigé.

Cela dit, cette diversité de moyens et de types de gouvernance aide à comprendre davantage qu'il n'existe pas une seule façon de réguler les SCL. La prise de conscience des impacts des autres cadres réglementaires peut aider la compréhension des risques que pourrait avoir l'Approche au Canada.

2.4 Objectifs et question de recherche

Dans le cadre de ce travail dirigé, j'effectuerai, en premier lieu, une analyse de l'Approche afin d'observer les risques que cette dernière pourrait avoir sur les environnements numériques au Canada, les communications et les vies sociales et politiques des Canadiens et Canadiennes.

En second lieu, je présenterai une synthèse de mes résultats d'analyse sous la forme d'une simulation d'un mémoire de consultation publique. Cette section du travail dirigé sera une réponse au gouvernement du Canada pour offrir ma perspective sur le sujet et des recommandations comme si je participais activement à la consultation publique qui a eu lieu durant l'été 2021. En présentant mes résultats d'analyse dans le format d'un mémoire de consultation, j'espère pouvoir utiliser ce chapitre dans une future consultation publique.

Ma question de recherche principale est la suivante : Quels sont les risques, les enjeux d'équité et de faisabilité que l'approche proposée par le gouvernement du Canada pour lutter contre le contenu préjudiciable en ligne pourrait avoir sur les environnements numériques au Canada, les communications et la vie sociale et politique des Canadiens et Canadiennes ?

3 MÉTHODOLOGIE

Dans ce chapitre, je présenterai l'analyse de politique publique, le cadre d'analyse que j'utiliserai et la technique de collecte de données. Comme mentionné précédemment, je commencerai par analyser l'Approche en utilisant un cadre d'analyse basé sur l'analyse de politiques publiques. Je présente aussi, dans ce chapitre, mon corpus de textes et quelques considérations éthiques.

3.1 L'analyse de politiques publiques

Devenue de plus en plus utilisée pour comprendre l'État par ses actions (Muller, 2000, p. 190), l'analyse de politiques publiques consiste à évaluer les enjeux et les impacts d'une proposition de politique sur différents acteurs. Ce type d'analyse permet de cerner les problèmes et aide les décideurs lorsqu'ils « […] ont l'obligation de rendre des comptes […] » (Morestin et le CCNPP, 2012, p. 2). Cependant, malgré le temps et les ressources investis dans des analyses de politiques publiques, plusieurs études empiriques démontrent que les décideurs les utilisent rarement pour l'amélioration des politiques (Shulock, 1999, p. 226). Malgré ce paradoxe, l'analyse de politiques publiques fait partie du processus démocratique puisqu'elle permet d'expliquer l'information et d'informer les citoyens. Plutôt que de voir ce type d'analyse comme un outil d'aide à la prise de décision, il faudrait le comprendre comme un outil qui permet d'avoir un discours public informé (Shulock, 1999, p. 241).

D'abord, certaines analyses sont dites descriptives ou *ex post* signifiant qu'elles s'effectuent une fois la politique en vigueur (Patton *et al*., 2016, p. 22 ; Bozio, 2018, p. 30). Elles observent les impacts et évaluent la performance de la politique. Ensuite, d'autres analyses de politiques sont dites prospectives ou *ex ante* (Patton *et al*., 2016, p. 23 ; Bozio, 2018, p. 28). Ces analyses se concentrent sur les résultats potentiels de la politique publique proposée (Patton *et al*., 2016, p. 23). Le troisième type d'analyses de politiques est une analyse prescriptive signifiant qu'elle offre des recommandations (Patton *et al*. 2016, p. 23). Ce travail sera principalement une analyse prospective ou *ex ante* puisque la politique n'est pas encore mise en place. Je tente d'informer les décideurs des risques que l'Approche pourrait avoir sur les Canadiens et Canadiennes et sur les environnements numériques. Le dernier chapitre aura toutefois une dimension d'analyse prescriptive puisque j'émettrai des recommandations.

Les analyses de politiques publiques peuvent aussi avoir divers objectifs. Elles peuvent être utilisées pour *déterminer si des propositions sont adéquates* avant même de devenir des projets de loi (Bozio, 2018, p. 28 ; Morestin et le CCNPP, 2012, p. 1). Une analyse peut aussi servir *à construire un argumentaire pour défendre une position* assumée face à une politique ou peut simplement « […] fournir au décideur les éléments d'information requis pour une décision éclairée. » (Morestin et le CCNPP, 2012, p. 1) Enfin, elle peut avoir un *objectif de comparaison avec d'autres politiques* (Morestin et le CCNPP, 2012, p. 2).

Au-delà de l'objectif, il n'existe pas un modèle qui s'applique à tous les cas (Patton *et al.*, 2016, p. 40 ; Howlett et Lindquist, 2004, p. 226). Patton *et al*. (2016) soutiennent qu'il existe certains éléments utiles, peu importe la politique étudiée ou le domaine (p. 3). Par exemple, ces auteurs proposent de suivre un modèle basé sur le processus rationnel de prise de décision, intitulé la méthode d'analyse de base, qui consiste à définir le problème, établir des critères d'évaluation, identifier des politiques alternatives, évaluer les politiques alternatives, démontrer les avantages et désavantages de chaque politique et finalement d'évaluer la politique une fois en vigueur (Patton *et al.*, 2016, p. 44-53). Le but de ce modèle est de fournir une analyse de base qui a des objectifs plus pratiques comme informer les décideurs suffisamment pour éviter des erreurs majeures (Patton *et al.*, 2016, p. 4).

D'un autre côté, certains proposent des cadres d'analyses qui varient un peu : définir le problème, collecter et décrire les éléments décisifs, analyser les éléments décisifs et formuler des recommandations ou décrire les alternatives possibles (Milovanovitch, 2018). La différence majeure entre le modèle de Patton *et al*. (2016) et celui de Milovanovitch (2018) est l'introduction de politiques alternatives au sein même de l'analyse.

Enfin, certains proposent une méthode clé en main d'analyse de politiques. Par exemple, le CCNPP divise l'analyse en six dimensions : l'efficacité, les effets non recherchés, l'équité, les coûts, la faisabilité et l'acceptabilité (Morestin et le CCNPP, 2012, p. 2). Dans ce cas, il suffit d'analyser la politique publique selon ces six critères d'évaluation. En bref, il n'y a pas un modèle parfait qui peut être appliqué à toutes les analyses de politiques publiques. L'analyste doit être en mesure de

faire des choix et adapter son modèle pour répondre le plus adéquatement aux objectifs. Je présente, dans le tableau 1, une synthèse de cette section.

Tableau 1 Synthèse des différents modèles et objectifs d'analyses des politiques publiques

Types	Objectifs
Ex post	• Observer les impacts ; • Évaluer la performance de la politique.
Ex ante	• Tente de déterminer les résultats potentiels de la politique avant qu'elle soit mise en place.
Prescriptif	• Émettre des recommandations à la suite de l'analyse.

3.2 Définition du cadre d'analyse

Dans ce travail, je m'inspirerai des différents cadres d'analyse de politiques publiques proposés. Comme mentionné ci-dessus, il n'y a pas un modèle clé en main d'analyse de politique publique. Chaque analyste doit faire des choix éclairés sur les éléments analysés. Par souci de temps et de ressources, je devrai limiter ce processus d'analyse afin de me concentrer sur les grandes sections de l'Approche comme définies dans le *Guide de discussion* et le *Document technique* : les entités réglementées, les catégories de contenu préjudiciable réglementées, les nouvelles règles et obligations et les nouveaux organismes de réglementation. J'ai choisi de laisser tomber l'analyse des modifications du cadre législatif canadien existant puisque je ne voulais pas m'attarder sur des formalités juridiques.

Les quatre axes principaux de mon analyse sont les suivants :

1. **Les entités réglementées :** L'Approche régule les « services de communication en ligne » (SCL) qui ont des activités au Canada. Le gouvernement du Canada définit les SCL comme étant des services qui permettent aux usagers de communiquer par Internet (*Document technique*). De cette Approche sont exclus, notamment, les communications privées, les fournisseurs de télécommunications et les SCL cryptées (*Guide de discussion*).
2. **Les catégories de contenus préjudiciables réglementées :** L'Approche vise à réglementer cinq catégories de contenus préjudiciables, soit le contenu terroriste, le contenu incitant à la violence, les discours haineux, le partage non consensuel d'images intimes et le contenu d'exploitation sexuelle d'enfants (*Guide de discussion* et *Document technique*). Les

définitions de chaque catégorie sont basées sur le Code criminel du Canada et seront présentées dans le chapitre d'analyse.

3. **Nouvelles règles et obligations :** De nouvelles règles et obligations seront mises en place comme un délai de retrait en 24 h, des mécanismes de signalement et d'appel robustes et des rapports de transparence (*Document technique*). Ces règles et obligations devront être respectées par les entités réglementées sous peine de sanctions.
4. **Nouveaux organismes de réglementation :** L'Approche prévoit de nouveaux organismes de réglementation pour assurer l'application des règles, offrir un processus d'appel pour les Canadiens et Canadiennes et prodiguer des conseils d'experts. Les nouveaux organismes sont le Commissaire à la sécurité numérique, le Conseil de recours en matière numérique du Canada et le Comité consultatif d'experts (*Guide de discussion et Document technique*). Leurs compositions et mandats seront détaillés dans le chapitre d'analyse.

Chaque axe sera ensuite analysé par trois critères d'évaluation : les risques, l'équité et la faisabilité. Le tableau 2 présente une représentation visuelle du cadre d'analyse.

Tableau 2 Cadre d'analyse de l'Approche

	Entités réglementées	Catégories de contenus préjudiciables réglementées	Nouvelles règles et obligations	Nouveaux organismes de réglementation
Les risques				
L'équité				
La faisabilité				

3.2.1 Les risques

L'Approche ne s'est pas rendue au stade de projet de loi officiel. Pour cette raison, j'analyserai les risques de cette dernière sur les environnements numériques, sur les droits et libertés des Canadiens et Canadiennes et sur les communications, vies sociales et vies politiques des usagers. Le risque peut être défini de deux façons : compris comme un danger potentiel (Gellert, 2018, p. 280) et compris comme l'évaluation de futures possibilités qui s'effectuent en prévoyant les événements futurs autant positifs que négatifs, et en utilisant cette analyse pour prendre des décisions (Gellert,

2018, p. 280). Ce critère d'analyse me permettra d'anticiper les impacts et conséquences possibles de l'Approche.

3.2.2 L'équité

Il est important d'analyser la dimension de l'équité, car il « […] s'agit de voir si la politique analysée produit des effets différents dans divers groupes […], ou encore si elle risque de provoquer, d'augmenter ou de corriger des inégalités […] » (Morestin et le CCNPP, 2012, p. 4). Au-delà d'analyser l'efficacité d'une politique publique, il est primordial de prendre en considération comment celle-ci impactera les différents groupes de la société. Une politique publique efficace pour une partie de la population peut s'avérer défavorable et néfaste pour d'autres.

3.2.3 La faisabilité

Ce critère d'évaluation consiste à analyser les ressources disponibles, d'évaluer si la politique peut entrer en application selon le cadre législatif déjà en place et si elle s'insère dans des processus administratifs déjà existants (Morestin et le CCNPP, 2012, p. 5). Je me concentrerai davantage sur la faisabilité au niveau technologique et des ressources humaines en analysant les outils utilisés par les entités réglementées pour savoir s'ils peuvent répondre adéquatement et dans les délais demandés aux règles et obligations et en évaluant si les ressources humaines sont suffisantes pour réaliser l'ampleur du mandat.

3.3 Définition de la structure de la simulation du mémoire de consultation publique

Lorsque le gouvernement du Canada ouvre une consultation publique, il n'y a pas de structure ou modèle exigé pour soumettre une réponse. En regardant quelques mémoires de la consultation publique, j'ai réalisé que chacun utilise la structure qui leur convient le mieux. La plupart regroupent différentes critiques sous des thématiques plus générales et plusieurs proposent des recommandations (Khoo *et al*., 2021 ; Geist, 2021 ; OpenMedia, 2021 ; Access Now, 2021).

Pour la structure de ma simulation de mémoire de consultation publique, j'ai décidé de m'inspirer du mémoire d'OpenMedia (2021), car je trouvais leur structure facile à suivre et très informative. La structure que j'utiliserai consistera à faire ressortir de mon analyse cinq grandes problématiques

de l'Approche. Ensuite, pour chaque problématique, je présenterai des constats généraux et proposerai des recommandations basées sur les différents documents lus pour l'analyse.

3.4 Corpus de textes

Mon analyse se rapproche davantage de l'étude descriptive qualitative (Fortin et Gagnon, 2016, p.199) qui consiste à décrire en profondeur un phénomène ou événement peu connu ou peu compris. Plusieurs techniques de collecte de données sont employées comme l'entrevue, l'observation et l'examen de documents (Sandelowski, 2000, p. 338 ; Fortin et Gagnon, 2016, p. 199). Dans le cadre de mon analyse, j'utiliserai l'examen des documents (Mouketou, 2021, p. 21). Cette technique me permettra de rassembler les connaissances sur les différents axes ciblés dans mon cadre d'analyse en consultant différents types de documentation.

L'information présentée dans cette analyse provient directement de mon corpus de textes principal, soit :

- Le *Guide de discussion* (Ministère du Patrimoine canadien, 2021a) qui présente le contexte, les objectifs, les motivations et les grandes lignes de l'Approche. Il s'agit d'un document relativement court qui passe très rapidement sur les différents éléments de l'Approche pour tout simplement donner une idée générale de la proposition du gouvernement.
- Le *Document technique* (Ministère du Patrimoine canadien, 2021b) qui reprend les différents modules et sous-sections de l'Approche pour expliquer plus en détail le fonctionnement, les attentes, les changements et les nouvelles obligations. Ce document permet de comprendre concrètement comment les différents éléments de l'Approche seront mis en place.
- Le *Rapport synthèse* (« Ce que nous avons entendu : Approche proposée du gouvernement pour s'attaquer au contenu préjudiciable en ligne ») (Ministère du Patrimoine canadien, 2022a) qui consiste à être un résumé des réponses reçues lors de la consultation publique. Ces réponses n'ont pas été publiées dans leur entièreté pour des raisons d'informations confidentielles, mais le gouvernement à décider de plutôt souligner les préoccupations qui avaient été soulevés le plus souvent dans les soumissions.

À titre de rappel, pour des fins de lisibilité et de concision, je référerai à ces différents documents en les nommant directement, plutôt qu'en citant leurs références : *Guide de discussion*, *Document technique* et *Rapport synthèse*.

Au niveau de mon analyse, j'utiliserai un complément de sources diverses pour appuyer mes propos et observations. Ce corpus de textes secondaire comprend de la littérature scientifique, de la littérature secondaire (articles de presses et articles de journaux spécialisés), des rapports (gouvernementaux et d'organismes) et des textes faisant référence à des cadres législatifs et réglementaires proposés dans d'autres juridictions. Ces sources me permettront de centraliser l'information connue sur le sujet et de donner une légitimité à mes propos.

3.5 Considérations éthiques

Au niveau des enjeux éthiques, je ne requiers pas de certificat éthique pour ma collecte de données puisque j'utilise uniquement des documents et données accessibles publiquement. Cependant, je tiens à mentionner que je suis consciente que mon analyse est teintée par mes expériences culturelles, mes valeurs et mes normes. Lorsque des recherches au sujet d'Internet sont effectuées, il est important de prendre en considération que plusieurs dimensions culturelles sont en jeu (Franzke *et al*., 2020, p. 15). Afin de minimiser l'impact que mes biais pourraient avoir sur les éléments mis de l'avant dans mon analyse, je tente de prendre une posture neutre qui représente le discours plus général des valeurs des Canadiens et Canadiennes.

4 ANALYSE

Cette section détaillera les différents axes principaux de l'Approche et les analysera à l'aide de trois critères : les risques, l'équité et la faisabilité.

4.1 Entités réglementées

Le concept de service de communication en ligne (SCL) est défini comme « […] des services accessibles à partir du Canada, qui ont pour objet principal de permettre à un utilisateur de ces services de communiquer par Internet. » (*Document technique*, art. 2) Soulignons que l'Approche exclut notamment les services « qui ne se qualifient pas comme des SCL » (par exemple, les applications d'entraînement ou les sites de voyage), les communications privées, les fournisseurs de services de télécommunications et les services cryptés (*Guide de discussion*). Notamment, les services qui ne seront pas assujettis à la réglementation sont, notamment, Facebook Messenger, la messagerie privée d'Instagram, Whatsapp, Telegram, Signal, WeChat, Discord, le Metaverse, les jeux vidéo et plusieurs autres. Parmi les entités nommées, plusieurs sont considérés comme des services de communication privée. À titre de comparaison, au Royaume-Uni, dans le Online Safety Bill, il est prévu de réguler, entre autres, les applications de messageries privées, les jeux en ligne, les fournisseurs de services infonuagiques et les moteurs de recherche (Department for Digital, Culture, Media & Sport, 2022).

Pendant plusieurs années, les communications privées et cryptées ont échappé aux cadres réglementaires puisqu'une tension existe entre la volonté de minimiser les contenus préjudiciables et les valeurs démocratiques comme la liberté d'expression et le droit à la vie privée (Andrey *et al*., 2021, p. 6). Le cryptage, par exemple, a longtemps été utilisé par des communautés marginalisées pour organiser des mouvements de résistance et avec l'arrivée de la technologie numérique, le cryptage était vu comme une opportunité pour les usagers de protéger leur vie privée et garder leur anonymat dans ce monde de plus en plus connecté (Myers West, 2018, p. 10). Dans les dernières années, le cryptage est devenu essentiel à la sécurité des sonneurs d'alertes, des journalistes et des défendeurs des droits humains (Myers West, 2018, p. 11). Rapidement, le cryptage a été perçu comme une nécessité pour la liberté d'expression dans l'ère des technologies numériques à un point tel que le Conseil des droits de l'homme des Nations Unies demande aux États d'éviter à tout prix d'interdire l'utilisation du cryptage (Myers West, 2018, p. 11). En excluant cette catégorie de SCL,

le gouvernement du Canada s'assure de ne pas empiéter sur le droit à la vie privée, la liberté d'expression et le droit d'association des Canadiens et Canadiennes (tous des droits fondamentaux de la personne).

Toutefois, on se questionne tout de même à savoir où se trouve la ligne entre une communication publique et une communication privée. De plus, les services de communications privées et cryptées sont un terreau fertile pour la propagation de contenus préjudiciables (Andrey *et al*., 2021). Je présente, dans le tableau 6, mes questions d'analyses pour cette section.

Tableau 3 Questions d'analyse pour l'axe des entités réglementées

	Questions d'analyse pour l'axe des entités réglementées
Risques	• Quels sont les risques d'exclure les communications privées et cryptées de l'Approche ? • Quels sont les risques de migration de certains groupes extrémistes ou terroristes vers des entités non réglementées ?
Équité	• Est-ce qu'une approche universelle pourrait nuire au développement de plus petits SCL ?
Faisabilité	• Comment le gouvernement canadien prévoit-il définir les entités réglementées et sur quelle base prévoit-il donner des exemptions ? • Est-ce possible de réguler les SCL privés ou cryptés sans enfreindre le droit à la vie privée des usagers ?

4.1.1 Risques

Pour commencer, il est primordial de souligner les risques liés aux entités qui ne seront pas assujetties à l'Approche, plus précisément, les services de communications privées et cryptées. À la suite d'un sondage effectué auprès de 2 500 résidents du Canada en mars 2020, 26 % des répondants affirmaient recevoir des propos haineux au moins une fois par mois (ou plus souvent pour les personnes de couleurs) par messagerie privée (Andrey *et al.*, 2021, p. 3). De plus, en 2020, à l'échelle internationale, il y aurait eu 20 millions de cas d'abus sexuels détectés sur Facebook. 99 % des cas auraient été détectés à l'aide d'outils d'automation et de ceux-ci, 70 % auraient été partagés dans les fonctionnalités de messagerie privée de Facebook et Instagram (Andrey *et al*., 2021, p. 19). Plusieurs de ces plateformes ont donc déjà des mécanismes en place pour filtrer le contenu non admissible selon leurs standards de communauté, et ce même dans les messages dits

privés. D'inclure ces SCL dans les entités réglementées ne demanderait pas nécessairement de mettre en place de nouvelles pratiques de modération de contenu. Cela donnerait plutôt un incitatif aux SCL qui ne le font pas de commencer à modérer le contenu qui circule dans les messageries privées. Cependant, cette solution n'est pas sans faute puisque cette tâche serait difficile pour les services de communication cryptés. L'utilisation du chiffrement permet de s'assurer qu'aucun intermédiaire n'ait accès au contenu du message envoyé (Andrey *et al*., 2021, p. 12). Seule l'entité qui a la clé de déchiffrement pourra y avoir accès. En conséquence, il resterait le risque qu'un plus grand volume de contenus préjudiciables se retrouve sur ce genre de plateforme afin d'éviter la modération et la surveillance des messages.

Ensuite, il est aussi important de noter un risque déjà observable dû au fait que les services de communications privées ou cryptées ne soient pas régulés : la migration des mouvements extrémistes vers ces plateformes (Dugal et Lozach, 2021). Les mouvements extrémistes souhaitent pouvoir s'exprimer librement sans subir de censure par les plateformes et à l'abri des yeux du gouvernement. Pour illustrer cet impact, prenons l'exemple de *Telegram*. *Telegram* est une application de messagerie privée. La fonctionnalité de clavardage secret est une messagerie cryptée, voulant dire que seules les personnes participant à la conversation peuvent avoir accès aux messages (aucun intermédiaire, même pas *Telegram*) et les messages se détruisent sans laisser de traces (Dugal et Lozach, 2021). L'application compte, maintenant, près de 500 millions d'utilisateurs (Dugal et Lozach, 2021).

À la suite de l'assaut au Capitole aux États-Unis, un mouvement de *déplatformisation* eut lieu. En d'autres mots, les médias sociaux comme Facebook, YouTube et Twitter ont procédé à bannir certains usagers ou comptes de leur plateforme (Dugal et Lozach, 2021). Cela dit, plusieurs usagers, principalement de mouvements extrémistes, ont migré vers *Telegram*. Ce phénomène a surtout été observé chez les adeptes de QAnon après qu'il y a eu plusieurs bannissements et suppressions de comptes (Dugal et Lozach, 2021). Ce qui attire les usagers vers l'application *Telegram* est la sécurité des données, l'éthique libertarienne et le fait qu'elle soit très peu modérée (Dugal et Lozach, 2021). Les mouvements conspirationnistes d'extrême droite et terroristes cherchent davantage le manque de modération, plutôt que l'anonymat (Dugal et Lozach, 2021). Durant l'année 2020, il a été estimé qu'environ 8 % des résidents canadiens auraient utilisé l'application

Telegram (Andrey *et al*., 2021, p. 33). Enfin, la limite de la taille des groupes est de 200 000 personnes (Andrey *et al*., 2021, p. 33). Est-ce encore considéré comme des messages privés ? Selon l'Approche, ce SCL ne serait pas assujetti aux cadres réglementaires, car il est considéré comme un service de communication privée. La grande visibilité qu'une personne peut avoir sur ce genre de SCL peut mener à des déraillements de partages de contenus préjudiciables comme de la propagande terroriste.

En bref, en excluant les services de communications privées, le gouvernement fait fi de la prédominance du contenu préjudiciable qui y circule. Plusieurs de ces services modèrent déjà les messages privés afin de bloquer le contenu non admissible. En les incluant dans l'Approche, le gouvernement du Canada donnerait un incitatif à tous les SCL de le faire. Enfin, le gouvernement prendrait aussi le risque de voir de plus en plus de mouvements extrémistes migrés vers des services de communications privées ou cryptées afin d'éviter la surveillance. En conséquence, au lieu que le contenu soit retiré d'Internet, l'Approche pousserait à tout simplement l'invisibiliser dans des plateformes non réglementées, où les contenus continueront à causer préjudices.

4.1.2 Équité

Au niveau de l'équité, avec la définition fournie dans le *Guide de discussion* et le *Document technique*, il ne semble pas y avoir d'enjeux de discrimination envers une communauté versus une autre. Cependant, le concept d'équité est compris comme l'augmentation d'inégalités de tous genres (Morestin et le CCNPP, 2012, p. 4). Dans cette ligne d'idée, la définition fournie, dans le *Guide de discussion* et le *Document technique*, semble placée tous les différents types de SCL dans le même panier. En d'autres mots, les grandes et petites entreprises seraient attendues de répondre aux mêmes exigences légales dans les mêmes délais. Le risque dans cette situation est de ne pas prendre en considération les différents modèles d'affaires, la taille, le volume de contenus et les ressources (humaines et monétaires) disponibles (Ministère du Patrimoine canadien, 2022c ; Caplan, 2018, p. 26). Selon certains participants de la consultation publique, il aurait été mieux de réguler par rapport au niveau de risque d'hébergement ou d'exposition à du contenu préjudiciable de chaque SCL (*Rapport synthèse*). Ceci permettrait aux règles et obligations d'être plus ou moins strictes selon le niveau de risque (Ministère du Patrimoine canadien, 2022c). Il s'agirait d'une alternative permettant une flexibilité dans l'application du cadre réglementaire et prendrait en

considération les différences au sein des SCL. Cela éviterait aussi de nuire au développement de SCL alternatifs puisqu'ils ne seront pas freinés par le fardeau financier et technique engendré par l'Approche (*Rapport synthèse*) et donc, éviterait des inégalités au sein du marché.

4.1.3 Faisabilité

Au niveau de la faisabilité, il y a des enjeux d'avoir une définition trop restreinte qui inclus nombreuses exemptions. D'abord, il peut devenir difficile à savoir quel SCL est assujetti ou non au cadre réglementaire. De plus, une définition trop restreinte peut nuire à la pérennité de l'Approche considérant que les environnements numériques sont en constants changements et que les lieux d'hébergement des contenus préjudiciables évoluent rapidement (Ministère du Patrimoine canadien, 2022c). En ayant une définition vaste, l'Approche restera pertinente face l'émergence de nouvelles technologies.

Enfin, dans une approche alternative qui régulerait les communications privées et cryptées, il existe déjà des mécanismes pour minimiser le partage de contenu préjudiciable. Le fait d'instaurer un cadre réglementaire pour ces services obligerait toutes les plateformes à devoir fournir des efforts pour modérer le contenu préjudiciable en ligne et éviter que des nids de mouvements extrémistes se créent à l'abri des yeux de la loi. Ci-dessous, une liste, non exhaustive, de moyens utilisés par des services de communications privées pour réduire la propagation du contenu préjudiciable est présentée :

- Des mécanismes de signalement par les usagers (Andrey *et al*., 2021, p. 36) ;
- La détection automatique (par l'utilisation d'algorithmes) des comptes ou contenus préjudiciables (Andrey *et al*., 2021, p. 19).
- Une limite sur le nombre de fois qu'un message peut être transféré. Ceci est mis en place pour tenter de ralentir la propagation de contenu. Les plateformes utilisent cette technique principalement pour limiter la propagation de désinformation. Il ne s'agit pas du tout d'une mesure sans faute, car il est possible de s'organiser en groupe pour déjouer ce garde-fou (Andrey *et al*., 2021, p. 37).
- Une limite sur la taille des groupes de discussion. Par exemple, Instagram a une limite de 32 personnes alors que *Telegram* a une limite de 200 000 personnes par groupe de

discussion. Il y a une grande différence dans la taille du public et dans la visibilité du contenu aura (Andrey *et al*., 2021, p. 38).

- La mise en place d'outils de vérification des informations reçues. Par exemple, Whatsapp permet d'effectuer une recherche sur Google des messages reçus (Andrey *et al*., 2021, p. 37).

Tableau 4 Points saillants de l'analyse des entités réglementées

Risques	Équité	Faisabilité
Risque d'une prolifération de contenus préjudiciables sur les entités non réglementées. Risque de voir les groupes extrémistes migrer vers des SCL privés et cryptés.	Risques d'inégalités entre les petites et grandes entreprises de SCL au sein du marché.	Une définition trop restreinte des entités réglementées pourrait nuire à la pérennité de l'Approche. Plusieurs moyens existent pour ralentir la propagation de contenus préjudiciables sur les SCL privés ou cryptés sans enfreindre le droit à la vie privée des usagers comme limiter la taille des groupes de discussion ou limiter le nombre de partages d'un même contenu.

4.2 Catégories de contenu préjudiciable réglementées

Cinq catégories de contenus préjudiciables, soit les plus flagrantes selon le gouvernement, sont ciblées et définies en se basant sur le Code criminel : le contenu terroriste, le contenu incitant à la violence, les discours haineux, le partage non consensuel d'images intimes et le contenu d'exploitation sexuelle d'enfants en ligne. Le tableau 8, ci-dessous, présente les définitions des différentes catégories de contenus préjudiciables.

Pour commencer, je me questionne sur la décision du gouvernement du Canada de cibler uniquement ces cinq catégories de contenus préjudiciables. En étudiant les cadres réglementaires à l'international, je ne peux m'empêcher de remarquer que chacun vise différentes catégories de contenus. De plus, le Canada semble être l'approche qui cible le moins de contenus parmi celles présentées dans la problématique de ce travail. Contrairement au Royaume-Uni, le Canada ne fait pas de distinction entre le contenu préjudiciable illégal et le « contenu préjudiciable, mais légal » (*harmful but legal*). Le « contenu préjudiciable, mais légal » est, entre autres, le harcèlement,

l'automutilation et les discours encourageant les troubles alimentaires (Department for Digital, Culture, Media & Sport et Dorries, 2022). En Australie, le Safety Online Act cible la cyberintimidation envers les enfants comme une catégorie de contenus préjudiciables (The Parliament of Australia, 2021). Enfin, la désinformation est aussi incluse dans le cadre réglementaire du Royaume-Uni (Department for Digital, Culture, Media & Sport, 2022).

Tableau 5 Définitions des cinq catégories de contenu préjudiciable

Types de contenus préjudiciables	Définitions des catégories
Le contenu terroriste	Le contenu terroriste est défini comme du « […] contenu qui encourage activement le terrorisme et qui est susceptible d'entraîner du terrorisme. » (*Document technique*, art. 8)
Le contenu incitant à la violence	Le contenu incitant à la violence fait référence à tout contenu qui encourage la violence ou qui pourrait mener à de la violence (*Document technique*, art. 8).
Les discours haineux	Selon le *Document technique*, les discours haineux feront référence à la définition dans la Loi canadienne sur les droits de la personne. En mars 2022, le projet de loi C-261 a été déposé pour apporter des modifications au Code criminel et à la Loi canadienne sur les droits de la personne. Ces modifications concernent les discours haineux et proposent la définition suivante : « […] discours haineux s'entend du contenu d'une communication qui, sur le fondement d'un motif de distinction illicite, exprime de la détestation à l'égard d'un individu ou d'un groupe d'individus ou qui manifeste de la diffamation à leur égard. » (Vuong, 2022, art. 13[9])
Le partage non consensuel d'images intimes	Ce type de contenu réfère à des cas où la personne représentée dans le contenu n'a pas donné son consentement ou s'il n'est pas possible de déterminer si consentement a été donné (*Document technique*, art. 8).
Le contenu d'exploitation sexuelle d'enfants en ligne	Cette catégorie fait référence à tout matériel qui a un lien (direct ou indirect) avec l'exploitation sexuelle des enfants et qui est préjudiciable pour les victimes (*Document technique*, art. 8).

Le Ministre du Patrimoine canadien en 2021, Steven Guilbeault, affirmait que le gouvernement du Canada ne prévoit pas réguler les « contenus préjudiciables, mais légaux » (*harmful but legal*), car ceux-ci, malgré leur caractère blessant et horrible, ne relèvent pas de la définition légale de discours haineux (L'avenir nous appartient, s.d.). Cependant, l'Approche ne s'attaque pas seulement aux discours haineux. Plusieurs « contenus préjudiciables, mais légaux », comme le harcèlement, l'automutilation et la désinformation, pourraient être inclus dans les préjudices afin

d'éviter de fermer les yeux sur ces contenus problématiques. Le tableau 9 présente mes questions d'analyses pour cette section.

Tableau 6 Questions d'analyse pour l'axe des catégories de contenu préjudiciable réglementées

	Questions d'analyse pour l'axe des catégories de contenu préjudiciable réglementées
Risques	• Comment le gouvernement du Canada peut-il définir ce qui constitue un préjudice alors qu'il n'y a pas consensus ? • Quels sont les risques d'avoir un éventail de catégories de contenus préjudiciables trop diversifié ? • Quels sont les risques de déterminer des catégories de contenus préjudiciables ?
Équité	• Quels sont les risques d'un manque de compréhension du contexte local lors de la modération de contenus préjudiciables pour les communautés marginalisées ? • Quels sont les risques que des contenus préjudiciables tombent sous les catégories de discours protégés ?
Faisabilité	• Quels sont les risques d'avoir des définitions précises et fixes de contenus préjudiciables en ligne ? • Est-ce que les membres des SCL, du Conseil de recours en matière numérique du Canada et le Comité d'experts auront l'expertise nécessaire pour prendre des décisions sur l'éventail diversifié de catégories de contenus préjudiciables ?

4.2.1 Risques

Avant d'entamer l'analyse des risques, il est important de faire un bref survol de la situation actuelle. Présentement, ce sont des acteurs privés, soit les entreprises de SCL, qui décident ce qui peut ou ne peut pas être publié ou hébergé sur leurs plateformes ou sites Web (DeNardis et Hackl, 2015, p.766 ; Mouketou, 2021, p. 12). Les SCL ont souvent été considérés neutres en se déresponsabilisant face aux contenus publiés sur leurs plateformes (DeNardis et Hackl, 2015, p. 766). Cependant, en pratique, les SCL définissent, à travers les standards de communauté et les termes et conditions d'utilisation, ce qu'ils considèrent comme du contenu non admissible (Gerrard and Thornham, 2020, p. 1276). L'Approche suggère plutôt d'imposer la loi canadienne aux acteurs privés afin que le concept de contenu préjudiciable soit défini par les lois officielles plutôt que les intérêts financiers des acteurs privés. Cependant, certaines questions demeurent : le gouvernement devrait-il réglementer et catégoriser les contenus préjudiciables ? Pourquoi avoir seulement ciblé

cinq catégories de contenus préjudiciables ? Pouvons-nous vraiment évaluer le préjudice causé par le contenu en ligne ?

À la suite d'un survol rapide de ce qui se fait dans d'autres pays démocratiques, comme l'Allemagne, l'Australie, la France et le Royaume-Uni, il est possible de remarquer qu'il ne semble pas y avoir un consensus sur ce qui devrait être intégré dans les catégories de contenus préjudiciables. Par exemple, la désinformation est régulée dans les cadres réglementaires de la France et du Royaume-Uni (République française, 2020 ; Department for Digital, Culture, Media & Sport, 2022), mais pas dans l'Approche du Canada. Pour ce qui est de la désinformation, il a été démontré qu'elle est souvent la source d'incitation à la violence, d'ingérence électorale, de radicalisation et d'abus sexuels (Andrey *et al*., 2021, p. 15 à 19). Ensuite, au Royaume-Uni, le cadre réglementaire prévoit cibler le « contenu préjudiciable, mais légal » (*harmful but legal content*) (Department for Digital, Culture, Media & Sport, 2022). Entre autres, ceci inclurait du contenu qui fait la promotion de l'automutilation, du harcèlement et de l'intimidation (Department for Digital, Culture, Media & Sport, 2022). Chaque gouvernement semble déterminer les préjudices selon les normes et valeurs de leur société. Cela rend d'autant plus difficile de savoir où tracer la ligne pour ce qui est préjudiciable ou non. Le Canada a choisi d'utiliser les définitions du Code criminel plutôt que tenter de définir de nouveaux préjudices (*Document technique*).

Ensuite, les opinions semblent être partagées entre le risque d'avoir un éventail de catégories de contenus préjudiciables trop diversifiés ou l'opinion qu'il faudrait élargir davantage les catégories de préjudices. D'un côté, parmi les réponses reçues lors de la consultation publique, plusieurs ont exprimé que les catégories de contenus préjudiciables régulées représentent un éventail trop diversifié pour être adressé dans un seul et même cadre réglementaire (*Rapport synthèse* ; Geist, 2021, p. 2 ; Khoo *et al*., 2021, §17). Sur le plan légal, chaque catégorie de contenus serait traitée sur des bases constitutionnelles, factuelles, pratiques et éthiques entièrement différentes (Khoo *et al*., 2021, §17). De plus, les contenus en lien avec des formes de discours (discours haineux, terroriste ou incitant à la violence) risquent de tomber sous une des catégories de discours protégés par l'alinéa 2b) de la Charte des droits et libertés du Canada, rendant la tâche de modération d'autant plus complexe (Khoo *et al*., 2021, §19). Par exemple, parmi les discours protégés au Canada, on retrouve l'expression artistique, la satire, la critique et la réappropriation de certains

termes préjudiciables par les groupes visés par ces termes (Khoo *et al*., 2021, §19). Les discours haineux, le contenu terroriste ou incitant à la violence ont tendance à modifier leurs codes pour pouvoir justement tomber dans ces zones grises de discours protégés (Badouard, 2021, p. 100). Il est important d'avoir des équipes de modération suffisamment expertes pour détecter ces changements par risque sinon de remettre en question l'efficacité de l'Approche (Geist, 2021, p. 2).

De l'autre côté, nombreux experts mentionnent que le risque de déterminer des catégories de contenus préjudiciables est trop restrictif par rapport à la panoplie de préjudices qui peuvent survenir en ligne (*Rapport synthèse* ; Ministère du Patrimoine canadien, 2022d). Puisque le préjudice est subjectif à chacun, « […] il est impossible de prévoir tous les préjudices. » (Ministère du Patrimoine canadien, 2022 d, §22) Par subjectif, le préjudice est compris comme une expérience individuelle qui pourrait varier grandement d'une personne à une autre. Le contenu est performatif (Butler, 2004). En d'autres mots, certains termes utilisés peuvent sembler banals aux yeux d'une personne, mais être très destructeurs aux yeux d'une autre : « […] le 'contenu' même de certaines formes de discours ne peut être compris qu'à travers l'action que ces discours accomplissent [performs]. » (Butler, 2004, p. 108) Cela dit, tenter de fournir une liste détaillée de contenus considérés préjudiciables ou même se limiter à simplement cinq catégories ciblées peut mener à causer l'oubli ou à ignorer d'autres contenus qui pourraient constituer des préjudices pour certains usagers. Considérant la subjectivité du préjudice, il y a peu d'information sur quelles catégories de contenus devrait être priorisée dans un cadre réglementaire et législatif. Une solution proposée pour tenter de mitiger ce risque est de prévoir des définitions de contenus préjudiciables qui prendront en compte le caractère individuel et contextuel des préjudices tout en reconnaissant comment l'intersectionnalité peut impacter la perception du contenu (Ministère du Patrimoine canadien, 2022d).

Enfin, les définitions fournies dans le *Document technique* sont relativement vagues et difficilement déchiffrables. Selon le principe de légalité, utilisé en droit criminel, une loi devrait être écrite avec des termes clairs et précis (Grădinaru, 2018, p. 290). En pratique, toutefois, les lois sont souvent rédigées de façon vague afin de permettre une certaine flexibilité et il devient nécessaire de faire appel aux interprétations pour comprendre ce qui est permis (Grădinaru, 2018, p. 290). Considérant que l'Approche s'appuie fortement sur le signalement des contenus

préjudiciables par les usagers, il y a un risque que des définitions vagues et ambigües limitent le pouvoir des usagers à modérer leurs environnements numériques par peur d'être dans l'erreur ou par un manque de compréhension (Access Now, 2021, p. 2 ; Housefather, 2019, p. 21). Le gouvernement du Canada pourrait tenter de clarifier les définitions, soit dans la loi ou dans des documents connexes.

4.2.2 Équité

D'abord, l'expérience et le pouvoir performatif de certains contenus dépendent grandement du contexte et de l'interprétation de ceux-ci. Cela dit, des enjeux d'équité doivent être considérés lorsqu'il est question d'exiger aux SCL de modérer les contenus préjudiciables, plus particulièrement les discours haineux (Roberts, 2016, p.2). Plusieurs SCL ont un modèle d'affaire de modération de contenu commerciale (*Commercial content moderation*) qui nécessite l'utilisation d'énormes équipes de modérateurs humains pour examiner le contenu signalé (Roberts, 2016 ; Roberts, 2017 ; Roberts, 2018 ; Caplan, 2018). Ces équipes sont habituellement dispersées partout dans le monde et travaillent dans des conditions misérables dues aux salaires très bas, à l'obligation de rester secret sur leur emploi et aux dommages psychologiques causés par les contenus horribles auxquels elles sont soumises tous les jours (Roberts, 2016, p. 1). Les décisions de modération de contenu impliquent régulièrement des jugements complexes qui requièrent d'un modérateur de connaître et comprendre les normes, mœurs sociales et valeurs des lieux pour lesquels le contenu est destiné (Roberts, 2016, p. 1). En d'autres mots, les modérateurs doivent s'imaginer le public cible et incarner un ensemble de valeurs (autres que les siennes souvent) pour prendre une décision.

Toutefois, ce n'est pas toujours si facile, entre autres pour les décisions sur des discours haineux. Les discours haineux sont fortement dépendants du contexte local et des dynamiques de pouvoirs (Caplan, 2018, p. 13). Le modérateur doit alors comprendre les référents racistes, misogynes, homophobes, culturels et historiques d'une culture qui lui est souvent non familière (Roberts, 2016, p. 2), en plus de prendre en considération les intentions de la personne qui publie, le public ciblé et le type de discours émis (humour, ironie, expression artistique et autres) (Caplan, 2018, p. 13). Et tout cela se fait en quelques secondes (Caplan, 2018, p. 14). L'enjeu d'équité principal est que les modérateurs humains pourraient laisser des contenus préjudiciables exister sur le SCL, car ils

n'avaient pas les référents nécessaires pour identifier comme étant préjudiciables. Présentement, les SCL, qui utilisent la modération de contenu commerciale, n'ont pas l'équité et la justice sociale comme priorité dans la formation des équipes de modérateurs humains (Roberts, 2016, p. 9). En conséquence, ceci mène à une sous-représentation des communautés marginalisées, ce qui mène éventuellement à des biais dans les décisions de modération de contenu.

Ensuite, en Allemagne, afin de répondre aux exigences du NetzDG, les SCL ont mis sur pieds des équipes locales spécialisées dans les plaintes propres à ce cadre réglementaire (Heldt, 2019, p. 9). Toutefois, ces équipes reçoivent seulement les cas qui sont acheminés par les équipes de modérateurs plus larges. L'enjeux dans ce processus est le fait que pour qu'une équipe spécialisée dans le contexte local examine le contenu signalé, celui-ci doit avoir excité sur la plateforme, signifiant que les usagers l'ont vu, ont interagi avec le contenu et l'ont repartagé donnant toujours plus de visibilité à des contenus préjudiciables (Roberts, 2016, p. 4). Ce long processus pour arriver à une révision de contenu préjudiciable met encore une fois les communautés marginalisées à risque puisqu'il y a plus de chance que le contenu préjudiciable qui les vise reste en ligne due au manque de représentation.

Enfin, aux États-Unis principalement, mais le Canada n'y échappe pas, il y a une longue tradition de voir l'utilisation de codes racistes ou homophobes dans la culture populaire comme l'humour ou la satire (Roberts, 2016, p. 4). Ces contenus, même camouflés, peuvent néanmoins causer un préjudice aux personnes visées. Cela dit, selon la Charte des droits et libertés, l'humour et la satire tombent sous la catégorie des discours protégés signifiant que ces formes de préjudices pourraient rester en ligne sur les SCL (Khoo *et al*., 2021, §19).

4.2.3 Faisabilité

L'Approche vient définir ce qu'on entend par préjudice et donc vient fixer cette notion dans une période spécifique. Cette situation s'applique principalement par rapport aux discours haineux. Le fait de déterminer ce qui est considéré comme du discours haineux est un acte performatif qui découle des valeurs sociétales. Selon Gillespie (2020), « [c] alling something hate speech is not an act of classification, that is either accurate or mistaken. It is a social and performative assertion that something should be treated as hate speech, and by implication, about what hate speech is. » (p. 3)

En d'autres mots, la définition de discours haineux est en constante évolution puisqu'il s'agit d'un travail de négociation entre les différents groupes et différentes valeurs de la société. Toutefois, cette notion de discours haineux ne peut pas évoluer s'il y a toujours consensus. En bref, il existe le risque que l'Approche soit trop rigide pour s'adapter aux changements rapides et constants qui caractérisent les contenus préjudiciables.

De plus, il existe le risque que les membres des SCL, du Conseil de recours en matière numérique du Canada[8] et le Comité d'experts[9] n'aient pas l'expertise nécessaire pour juger tous les cas de cet éventail de contenus préjudiciables trop diversifiés (*Rapport synthèse*). Comme mentionné ci-dessus, les cinq différentes catégories de contenu préjudiciable doivent être évaluées sur des bases constitutionnelles et factuelles différentes (Khoo *et al*., 2021, §17). Par exemple, un cas de partage non consensuel d'images intimes ne serait pas jugé avec les mêmes critères qu'un contenu qui a rapport au discours, tel un discours haineux. Selon la Charte des droits et libertés, les questions relevant de l'expression doivent prendre en considération la nature et la gravité du préjudice causé avant de statuer sur un jugement (Khoo *et al*., 2021, §18). Il s'agit d'une analyse de cas complexe qui doit prendre en compte les éléments contextuels pour en comprendre l'impact réel. Afin de comprendre cette réalité, il est nécessaire d'avoir un certain niveau d'expertise, élément qui pourrait être difficile à atteindre pour prendre en compte cet éventail de préjudices dans un même cadre réglementaire.

Enfin, pour tenter de remédier à ce risque, il a été proposé d'utiliser une approche basée sur le risque (Ministère du Patrimoine canadien, 2022e). Concrètement, ceci signifie que les SCL auraient l'obligation d'effectuer des évaluations de risques et mettre en place les mesures nécessaires pour y répondre. Le gouvernement du Canada imposerait « […] des normes de rendement, au moyen de règlements et de lignes directrices, et d'évaluation de produits, au moyen de rapports de transparence et de vérifications […] » (Ministère du Patrimoine canadien, 2022e). L'idée derrière cette proposition est de miser davantage sur la notion de « devoir d'agir de manière responsable » ou le « devoir de diligence » afin de permettre une flexibilité aux SCL d'innover dans leurs moyens

[8] Voir la section 4.4 pour le Conseil de recours en matière numérique du Canada tel que présenté dans l'Approche.
[9] Voir aussi la section 4.4 pour le Comité d'experts tel que présenté dans l'Approche.

de modérer le contenu préjudiciable et d'avoir un cadre adaptable aux changements (Ministère du Patrimoine canadien, 2022e).

Tableau 7 Points saillants de l'analyse des catégories de contenu préjudiciable

Risques	Équité	Faisabilité
Risque d'avoir un éventail trop diversifié de catégories de contenus préjudiciables. Risque d'être trop restrictif dans les catégories de contenus préjudiciables ciblées.	Enjeux concernant les décisions de modération de contenu basées sur le contexte dû à un manque de représentation des communautés marginalisées au sein des SCL. Risque de voir des contenus préjudiciables être catégorisés comme discours protégés.	Risque de nuire à l'évolution de ce que la société considère ou non comme « contenu préjudiciable », surtout par rapport aux discours haineux. Risque que les SCL, le Conseil de recours en matière numérique du Canada et le Comité d'experts n'aient pas l'expertise nécessaire pour juger les contenus préjudiciables des différentes catégories ciblées.

4.3 Nouvelles règles et obligations

De nouvelles règles et obligations sont mises en place dans le cadre de cette Approche. D'abord, tout contenu signalé par les usagers doit être traité par le SCL et sera considéré comme préjudiciable jusqu'à preuve du contraire (*Document technique*, art. 11 [A]). Si le contenu est reconnu comme étant préjudiciable, le SCL doit le rendre inaccessible au Canada dans un délai de 24 h suivant le signalement initial (*Guide de discussion*). Le *Document technique* note que le « gouverneur en conseil » pourrait proposer des délais différents selon le type de contenu préjudiciable (art. 11 [A]). La figure 2 représente le traitement des signalements.

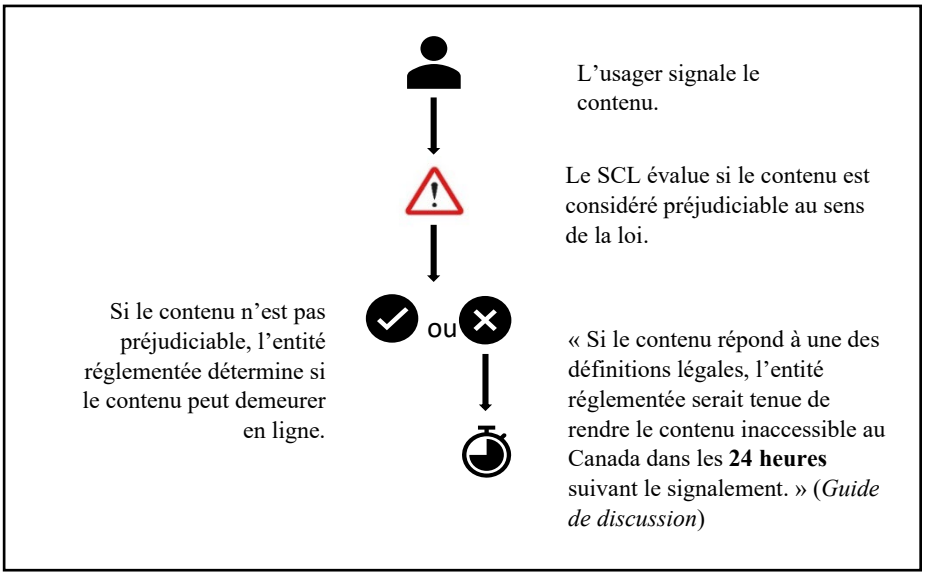

Figure 2 Traitement du contenu signalé

De plus, les entités réglementées devront « […] établir des systèmes robustes de signalement, de notification et d'appel, tant pour les auteurs de contenu que pour ceux qui signalent le contenu. » (*Guide de discussion*) Enfin, afin d'assurer plus de transparence, les entités réglementées auront l'obligation de publier, notamment, les données par rapport au volume et types de contenus traités au Canada, les changements apportés aux lignes directrices concernant le contenu admissible et des rapports sur l'impact des outils d'automation utilisés dans le processus de modération de contenu (*Guide de discussion* ; *Document technique*, art. 14).

L'élément qui semble le plus controversé est l'obligation de retrait en 24 h suivant le signalement initial. Une des craintes concernant le court délai est que les SCL procèdent à de la suppression excessive de contenu (*over-removal*). Ensuite, la nécessité d'avoir plus de transparence dans les processus de modération de contenu semble être une priorité autant pour les experts du domaine (Gorwa *et al*. 2020, p. 2 ; Roberts, 2018) que pour les différents gouvernements démocratiques comme l'Allemagne, le Royaume-Uni, la France et l'Australie requièrent des rapports de transparence mensuels ou annuels (Center for Democracy & Technology, 2017; Department for Digital, Culture, Media & Sport; République française, 2020 ; eSafety Commissioner, 2022).

Tableau 8 Questions d'analyse pour l'axe des nouvelles règles et obligations

	Questions d'analyse pour l'axe des nouvelles règles et obligations
Risques	• Quels sont les risques d'utiliser davantage d'outils d'automation ? • Quels sont les risques d'introduire un délai de 24 h suite au signalement initial pour rendre le contenu inaccessible au Canada ? • Quels sont les risques d'avoir une approche fortement basée sur les mécanismes de signalement des SCL ?
Équité	• Comment les communautés marginalisées seront-elles impactées par une utilisation accrue d'outils d'automation pour la modération de contenu préjudiciable ? • Est-ce que le délai de 24 h augmentera la surveillance des communautés marginalisées ?
Faisabilité	• Est-ce que le design du bouton de signalement peut avoir un impact sur l'efficacité de l'Approche ? • Est-ce que les moyens employés pour contourner les outils de modération de contenu automatisée peuvent nuire à l'efficacité de l'Approche ?

4.3.1 Risques

Premièrement, les nouvelles règles et obligations légales de l'Approche suggèrent l'utilisation d'outils d'automation afin de modérer le contenu préjudiciable en ligne. Le *Guide de discussion* et le *Document technique* mentionnent que les entités réglementées doivent faire tout dans leur pouvoir pour surveiller le contenu préjudiciable sur leurs services, incluant en utilisant des systèmes automatisés (*Guide de discussion* ; *Document technique*, art. 10). Avec la pression grandissante des gouvernements et des cadres réglementaires qui imposent des délais relativement courts pour le retrait de contenu, les SCL sont obligés d'utiliser de plus en plus des systèmes automatisés pour détecter le contenu préjudiciable (Gorwa *et al*., 2020, p. 2). Les outils d'automation sont perçus, en quelque sorte, comme la solution qui réglera tous les problèmes complexes de modération de contenu (Gorwa *et al*., 2020, p. 2). Le risque est alors d'obliger les SCL à s'appuyer principalement sur des outils d'automation et effacer les décisions humaines du processus de modération de contenu. Ce dernier est constamment un jeu d'équilibre entre prioriser le contexte ou prioriser la cohérence (Caplan, 2018).

Rappelons que selon Caplan (2018), il existe trois approches à la modération de contenu : l'approche communautaire, l'approche artisanale et l'approche industrielle. La modération de contenu sur les plateformes qui utilisent l'approche communautaire est faite par la communauté.

Ce sont des bénévoles qui s'assurent de l'acceptabilité des contenus selon les lignes directrices de la communauté virtuelle. Ensuite, l'approche artisanale priorise une approche de cas par cas. Il y a très peu d'utilisation d'outils d'automation et les signalements sont analysés en groupe avant de rendre une décision. Les deux premières approches sont très propices à prioriser le contexte local dans les processus décisionnels de modération de contenu et s'appuient principalement sur des décisions humaines plutôt que sur des systèmes automatisés pour la prise de décisions. Ensuite, l'approche industrielle, de son côté, utilise autant des outils d'automation que des modérateurs humains. Cette approche est souvent utilisée par les plus grandes plateformes et la cohérence sera priorisée. Puisque des milliers de personnes travaillent partout dans le monde pour modérer le contenu de ces plateformes, les règles de ce qui est accepté ou non doivent être simples, claires et faciles à appliquer.

Cela dit, comme mentionné auparavant, lorsque vient le temps de modérer les contenus préjudiciables, surtout des discours haineux, la compréhension du contexte local et des référents culturels est primordiale :

> « For a moderator to accurately assess whether content is hateful, they need to know the context of the content as it was made, including information about the individual making it, the target, and the environment, as well as linguistic or cultural clues they may not have access to […] » (Caplan, 2018, p. 13).

En conséquence, en prenant le risque d'obliger les SCL à s'appuyer principalement sur des systèmes automatisés, le gouvernement rend difficile la modération de contenu communautaire et artisanale, approche de modération alternative à celle industrielle. De plus, en priorisant une approche industrielle (ou commerciale), la cohérence des règles primera au détriment de la prise en considération des contextes locaux dans le processus décisionnel, ce qui peut poser problème au niveau des discours haineux (voir la section précédente).

De plus, la combinaison d'un grand volume de signalements et le court délai de retrait peut rendre difficile le traitement par des équipes de modérateurs humains. En conséquence, les SCL devront se doter de systèmes d'automation pour répondre aux exigences légales. Par exemple, le Network Enforcement Act (NetzDG), en Allemagne, oblige les plateformes à retirer certains contenus illégaux en 24 h et d'autres contenus préjudiciables dans un délai de 7 jours (Center for Democracy & Technology, 2017). En 2018, YouTube et Twitter mentionnaient, dans leurs rapports de

transparence, avoir chacun reçu plus de 200 000 signalements reliés au NetzDG (Tworek et Leerssen, 2019, p. 4). Ce nombre élevé de signalements représente une tâche colossale donc, les SCL s'appuient sur des outils d'automation pour filtrer une bonne partie des signalements.

Cependant, il est important de noter que les SCL sont conscients de l'incapacité des algorithmes à bien détecter les nuances dans les différents types de discours (Heldt, 2019, p. 9). Revenons à nouveau au cas du NetzDG en Allemagne. Pour faciliter l'application de la loi, les SCL utilisent un processus de modération à deux paliers. Une première grosse équipe, à l'échelle internationale, filtre le contenu qui ne respecte pas les lignes directrices de la plateforme. Ensuite, une plus petite équipe spécialisée en traitement des plaintes du NetzDG et qui parle allemand s'occupe des plaintes ayant un lien avec cette loi (Heldt, 2019, p. 9). On peut supposer que les SCL utiliseront un processus similaire si l'Approche était pour devenir une loi officielle. Toutefois, il ne faut pas complètement mettre de côté un risque d'une utilisation accrue des outils d'automation qui pourraient avoir des impacts importants sur l'équité, élément qui sera détaillé dans la sous-section suivante.

Deuxièmement, il y a un risque d'enfreindre la liberté d'expression et nuire à la documentation des cas de non-respect des droits humains puisque le court délai de 24 h peut mener à une suppression excessive de contenu de la part des SCL (Banchik, 2021, p. 1528 ; Baghdasaryan et Gullo, 2021 ; Article 19a, 2021). De plus en plus, des défendeurs des droits humains à travers le monde utilisent les médias sociaux pour documenter des preuves afin d'amener en justice certains dirigeants ou groupes extrémistes (Banchik, 2021, p. 1528). Toutefois, avec la pression des gouvernements de modérer plus sévèrement ce genre de contenu et les exigences de retrait, ces groupes d'activistes ont de plus en plus de difficulté à documenter ces preuves. Déjà, plusieurs sources disparaissent dû au fait que les SCL ne veulent pas prendre de chance d'enfreindre des lois ou de se faire critiquer par le public. Cependant, même après appel et révision des décisions, les contenus ou profils sont très rarement remis en ligne (Banchik, 2021, p. 1528).

Ensuite, ce délai assez court oblige les SCL à prendre des décisions très rapidement sur des cas relativement complexes. Le problème ne se situe pas dans les contenus clairement illégaux, mais plutôt dans ceux qui sont ambigus, dans une sorte de zone grise. La combinaison du délai de 24 h

et les sanctions monétaires imposantes mènent les SCL à choisir l'option sûre, soit celle de retirer le contenu afin d'éviter les sanctions. Le NetzDG, en Allemagne, aurait pu donner un aperçu des conséquences d'imposer un délai de retrait, mais les rapports de transparence ne fournissent pas de données sur cet aspect de la loi. Ils n'adressent pas le nombre de signalements erronés ou le nombre de cas de suppressions erronés (Tworek et Leerssen, 2019, p. 4). En conséquence, il est difficile d'émettre des conclusions sur le délai d'inaccessibilité en 24 h puisqu'il s'agit d'un élément pour lequel il y a très peu de données pour appuyer les hypothèses émises.

Enfin, l'Approche requiert les SCL à mettre en place des systèmes de signalements robustes afin de détecter et traiter les contenus préjudiciables en ligne. Ceci signifie que le gouvernement s'attend à ce que les usagers participent activement à la lutte contre les contenus préjudiciables. Les mécanismes de signalement sont habituellement reconnus pour être une bonne initiative d'intégrer les usagers dans le processus de modération de contenu : « Each of these mechanisms allows users to participate in how the platform content is organized, ranked, valued, and presented to others. » (Crawford et Gillespie, 2016, p. 411) Les usagers ont alors un rôle actif dans la modération du discours en ligne. Cependant, il existe le risque que ces systèmes de signalements subissent une ludification. En d'autres mots, les systèmes de signalements sont parfois utilisés pour faire des blagues, pour générer de la visibilité ou font l'objet d'actions coordonnées (Crawford et Gillespie, 2016, p. 420). Cela dit, la ludification des systèmes de signalement délégitimise la catégorisation d'offenses (Crawford et Gillespie, 2016, p. 420). À travers ces jeux de signalement, certains groupes marginalisés peuvent être ciblés par des actions coordonnées d'usagers qui souhaitent limiter leur participation aux discours publics. Sur les SCL, la visibilité est perçue comme un indicateur de légitimité et lorsqu'un contenu est signalé, celui-ci n'est plus visible (Crawford et Gillespie, 2016, p. 421). En bref, les mécanismes de signalement peuvent aider les SCL à détecter de nouvelles formes de contenus préjudiciables, surtout considérant que ceux-ci évoluent rapidement. Toutefois, il y a un risque important de voir ces systèmes transformés en jeu pour limiter la participation de certains dans la sphère publique.

4.3.2 Équité

D'abord, les standards de communauté, comme le nom laisse supposer, devraient habituellement refléter les standards, les attentes et l'opinion de la communauté face aux comportements

acceptables sur la plateforme. Cependant, ces derniers sont très rarement écrits en collaboration avec les membres de la communauté, mais plutôt rédigés par des employés du SCL et communiqués aux usagers. En réalité, l'exercice d'élaborer les standards de communauté est subjectif et reflète les biais et les visions du monde des décideurs au sein des SCL. Ces normes, valeurs et biais sont ensuite intégrés aux processus de modération de contenu et dans les outils d'automation, retirant leur « neutralité » (Gerrard and Thornham, 2020, p. 1271).

Ensuite, la plupart des SCL utilisent des algorithmes d'apprentissage automatique (*machine learning tools*) pour modérer le contenu sur leurs plateformes. L'apprentissage automatique utilise des données déjà existantes pour tenter de trouver une correspondance avec les contenus déjà catégorisés comme étant « non désirés » (Gerrard and Thornham, 2020, p. 1269; Gillespie, 2020). En d'autres mots, le contenu doit déjà avoir été classé comme préjudiciable pour que les outils d'apprentissage automatique puissent les détecter et prendre action. En conséquence, les bases de données utilisées figent dans une période précise ce qui est acceptable ou non. Elles n'ont pas la même flexibilité d'évolution que la communauté (Binns *et al*., 2017, p. 7). De plus, l'action de classifier ce qui est acceptable ou non est toujours un processus subjectif qui est influencé par les biais et la vision du monde des personnes qui déterminent les règles et qui codent les outils d'automation. Considérant qu'il y a un manque de représentation au sein des SCL, les biais historiques, racistes, homophobes, sexistes et xénophobes sont imbriqués dans les outils d'automation et reproduisent des préjudices envers certaines communautés.

Enfin, le délai de retrait en 24 h encourage la surveillance proactive du contenu qui circule en ligne (Access Now, 2021 ; Khoo *et al*., 2021, p. 4). Cette surveillance proactive désavantagera disproportionnellement les communautés marginalisées puisqu'elles sont déjà exposées à plus de surveillance étatique et commerciale dans le système de capitalisme de surveillance (Myers West, 2018, p. 11 ; Khoo *et al*., 2021, p. 4).

En bref, l'utilisation accrue des outils d'automation risque d'exacerber, dans les environnements en ligne, les biais historiques déjà existants dans la société dû aux valeurs, normes et vision du monde imbriquées dans les codes de ses systèmes automatisés. Au-delà des outils d'automation, la

surveillance proactive requise pour respecter le délai de 24 h impactera négativement les communautés marginalisées.

.

4.3.3 Faisabilité

Premièrement, le design du bouton de signalement peut avoir un grand impact sur l'application des nouvelles règles et obligations (Tworek et Leerssen, 2019 ; Heldt, 2019, p. 12). Considérant que les nouvelles obligations légales s'appliquent sur le contenu signalé, il est primordial que les usagers puissent facilement signaler le contenu préjudiciable. Retournons à nouveau à l'exemple du NetzDG en Allemagne. Un écart majeur existe entre le nombre de signalements NetzDG sur Facebook versus sur YouTube ou Twitter. D'un côté, YouTube et Twitter donnent la possibilité de signaler directement le contenu comme tombant sous la régulation du NetzDG (Tworek et Leerssen, 2019, p. 5). Le bouton est donc facilement accessible et ne demande pas d'efforts supplémentaires de la part des usagers. D'un autre côté, Facebook a rendu le bouton de signalement du NetzDG difficilement accessible ce qui expliquerait le chiffre considérablement plus bas de signalements (Tworek et Leerssen, 2019, p. 5). Une personne doit cliquer sur plusieurs liens et être redirigée sur une nouvelle page Web avant de pouvoir signaler le contenu sous la loi NetzDG (Tworek et Leerssen, 2019, p. 5 ; Heldt, 2019, p. 12). De plus, lorsqu'un usager fait ce genre de signalement sur Facebook, un message d'avertissement s'affiche informant les usagers du risque de sanctions légales dans des cas de faux signalements (Heldt, 2019, p. 12). D'un côté, ceci ralentit les trolls Internet et les chances de ludification des mécanismes de signalement. Toutefois, d'un autre côté, cet avertissement peut inhiber un usager de signaler par peur de subir des conséquences s'il est dans l'erreur (Heldt, 2019, p. 12). Ces choix de design ont un impact important sur l'application des obligations légales, car en rendant le signalement difficilement accessible aux usagers, plusieurs cas de contenus préjudiciables pourraient continuer à circuler sur les SCL au Canada, nuisant à l'efficacité de l'Approche.

Deuxièmement, l'Approche oblige des rapports de transparence de la part des SCL (*Guide de discussion* ; *Document technique*). En Allemagne, dans le cadre du NetzDG, les rapports de transparence ne donnent pas suffisamment d'information pour évaluer l'efficacité et les impacts de cette loi (Tworek et Leerssen, 2019, p. 8) et ceci pourrait se produire au Canada. Une solution proposée pour remédier à ce risque est d'obliger chaque SCL à avoir un répertoire des différents

cas signalés et des décisions prises par rapport à ces cas, un peu comme le système mis en place au Canada pour les publicités politiques sur Facebook (Tworek et Leerssen, 2019, p. 8). Au niveau de l'Approche, il est à noter que les SCL devront, notamment, évaluer l'impact des outils d'automation utilisés (*Document technique*). Il reste à savoir s'ils sont vraiment en mesure de fournir des rapports sur ces différentes questions.

Enfin, l'Approche n'est pas à l'abri des contournements des systèmes de modération de contenu (Gerrard, 2018 ; Caplan, 2018, p. 12). Prenons, par exemple, le contenu qui encourage les troubles alimentaires (Gerrard, 2018). Plusieurs SCL, comme Instagram, Tumblr et Pinterest, bloquent ce genre de contenu en modérant les mots-clics (*hashtags*) (Gerrard, 2018, p. 4494). En d'autres mots, aucun résultat n'apparaît lorsque les usagers cherchent des mots-clics qui ont été catégorisés comme pro troubles alimentaires. Les algorithmes sont aussi entraînés pour retirer le contenu qui a déjà été ciblé comme pro troubles alimentaires (Gerrard, 2018, p. 4495). Cependant, les usagers ont commencé à utiliser des codes spécifiques à la communauté pour signaler qu'ils sont pro troubles alimentaires afin que les algorithmes et les modérateurs humains ne comprennent pas qu'il s'agit de ce genre de contenu (Gerrard, 2018, p. 4500). En d'autres mots, ces méthodes de modération de contenu semblent être efficaces pour éviter que de nouveaux usagers trouvent le contenu préjudiciable. Toutefois, la communauté continuera de partager ce contenu (Gerrard, 2018, p. 4508). Quoique les troubles alimentaires ne seraient pas considérés comme du contenu préjudiciable par l'Approche, cet exemple démontre que les usagers finissent par trouver des manières de contourner la modération de contenu, diminuant l'efficacité de l'Approche.

Tableau 9 Points saillants de l'analyse des nouvelles règles et obligations

Risques	Équité	Faisabilité
Risque de prioriser les outils d'automation au détriment des décisions humaines et de la prise en compte des contextes locaux. Risque d'enfreindre la liberté d'expression due au délai de rendre le contenu inaccessible dans les 24 h suivant le signalement initial. Risque de ludification des systèmes de signalements.	Les outils d'automation reproduisent et exacerbent les biais racistes, historiques, homophobes et sexistes. Risque de surveillance proactive accrue.	Le design du bouton de signalement peut grandement impacter l'efficacité de l'Approche. Les contournements des mécanismes de modération de contenu peuvent nuire à la détection des contenus préjudiciables.

4.4 Nouveaux organismes de réglementation

L'Approche propose la mise en place de nouveaux organismes de réglementation. La *Commission canadienne de sécurité numérique* (Commission) est créée et elle inclut trois organismes de réglementation : le Commissaire à la sécurité numérique (le Commissaire), le Conseil de recours en matière numérique du Canada (le Conseil de recours) et le Comité consultatif d'experts (le Comité) (*Guide de discussion* ; *Document technique*). Le tableau 13 présente les rôles, mandats et compositions des nouveaux organismes de réglementation.

Toutes ordonnances de non-conformité devront être publiques. Toutefois, le commissaire peut choisir de diffuser ou non le nom du SCL. De plus, le commissaire doit s'assurer de garder privée l'identité du plaignant et de toutes personnes liées au contenu. Enfin, le Commissaire peut effectuer des inspections à n'importe quel moment (et non pas seulement à la suite de plaintes) (*Document technique*, art. 88). Un mandat sera nécessaire uniquement pour entrer dans une maison d'habitation.

Tableau 10 Rôles, mandats et compositions des nouveaux organismes de réglementation

Nouveaux organismes de réglementation	Rôles, mandats et compositions
Le Commissaire à la sécurité numérique	Le Commissaire a pour mandat d'administrer, superviser et appliquer les exigences législatives. De plus, il dirige des recherches et doit supporter les entités réglementées dans leur lutte contre les contenus préjudiciables (*Guide de discussion*). Enfin, le Commissaire se verra octroyer certains pouvoirs comme la réception de plaintes des usagers, le traitement des cas de non-collaboration, la recommandation des sanctions pour les cas de non-conformité et la possibilité de demander d'imposer des blocages sur certains SCL dans des cas de non-conformité en matière d'exploitation sexuelle des enfants et/ou du contenu terroriste (*Guide de discussion*). À noter que le Commissaire peut refuser une plainte, entre autres, si elle est jugée « […] futile, vexatoire, ou faite de mauvaise foi […] » (*Document technique*, art. 42).
Le Conseil de recours en matière numérique du Canada	Le Conseil de recours est une cour d'appel indépendante pour les décisions de modération de contenu prises par les entités réglementées (*Guide de discussion*). Les plaintes pourront être traitées par audience publique ou en huis clos selon la situation (*Document technique*, art. 59). Cet organisme rend des décisions contraignantes lorsque le contenu est considéré comme préjudiciable au sens de la loi. Il sera composé de trois (3) à cinq (5) membres (*Document technique*, art. 46) d'expertises diversifiées et devra soumettre des rapports publics afin d'assurer la transparence des activités.
Le Comité consultatif d'experts	Le Comité est mis en place afin de fournir des conseils d'experts au Commissaire et au Conseil de recours. Il est composé de maximum sept (7) membres à temps partiel, tous nommés par le ministre du Patrimoine canadien (*Document technique*, art. 71). Le gouvernement insiste que le Comité soit composé d'experts diversifiés et de membres de différents groupes de la société (*Guide de discussion*). Enfin, il ne sera pas impliqué dans le processus décisionnel ou dans l'application de la loi.

Dans le *Rapport synthèse*, plusieurs enjeux et inquiétudes ont été soulevés par les répondants. Entre autres, un des enjeux majeurs est l'idée de remettre à un organisme indépendant la tâche de prendre des décisions concernant la liberté d'expression. De plus, la taille des différents organismes a soulevé énormément de questionnement autant au niveau des répondants que du groupe d'experts qui travaillent sur l'amélioration de l'Approche.

Tableau 11 Questions d'analyse pour l'axe des nouveaux organismes de réglementation

	Questions d'analyse pour l'axe des nouveaux organismes de réglementation
Risques	• Est-ce que les nouveaux organismes de réglementation creusent un déficit démocratique ? • Quels sont les risques associés au huis clos ? • Quels sont les risques associés aux nouveaux pouvoirs d'inspection des organismes de réglementation ?
Équité	• Les nouveaux organismes de réglementation répondent-ils à l'objectif de représenter la diversité culturelle ?
Faisabilité	• Est-ce que le Conseil de recours sera en mesure de répondre à toutes les plaintes ?

4.4.1 Risques

Premièrement, les processus de création et sélection des membres pour faire partie des nouveaux organismes de réglementation creusent un déficit démocratique (*Rapport synthèse*). En d'autres mots, les personnes occupant des rôles dans ces organismes ne sont pas démocratiquement élues, mais plutôt nommées par le ministère du Patrimoine canadien (*Document technique*, art. 71). Cela dit, les personnes qui ont la responsabilité de prendre des décisions par rapport à la régulation du discours public en ligne ne seront pas nécessairement représentatives de la population canadienne. De plus, il y a très peu de garde-fous mis en place pour s'assurer de la redevabilité de ces organismes auprès des Canadiens et Canadiennes (*Guide de discussion* ; *Document technique* ; *Rapport synthèse*). Déjà, par le fait même de ne pas être élu par les citoyens, ces derniers n'ont pas l'obligation d'être redevables, mais au-delà de cela, le seul mécanisme de vérification du travail des nouveaux organismes de réglementation est l'obligation de publier un rapport annuel (*Document technique*, art. 76).

Le Commissaire et le Conseil de recours doivent publier, à la fin de l'année fiscale, un rapport qui comprend, entre autres, les plaintes reçues et les décisions prises par rapport à celles-ci (*Document technique*, art. 77). Le Commissaire devra aussi fournir de l'information par rapport aux inspections et signalements qui auront été effectués durant l'année fiscale (*Document technique*, art. 77c). Bien que les rapports annuels semblent être assez complets, réaliser cet exercice uniquement une fois par année n'est pas suffisant considérant que les environnements numériques et les contenus préjudiciables en ligne évoluent très rapidement (*Rapport synthèse* ; Internet Society

Canada Chapter, 2021, p. 17). Afin d'assurer un meilleur processus démocratique et de s'assurer que ces organismes mettent de l'avant les intérêts de la population canadienne, il serait nécessaire de pouvoir évaluer leur travail à une fréquence plus élevée afin de corriger le tir avant qu'une situation dégénère (*Rapport synthèse*). Enfin, il n'y a pas mention du nombre d'années qu'un membre peut siéger dans ces organismes, aucune mention du processus pour retirer un membre si ce dernier ou cette dernière ne répond pas aux exigences et aucune mention concernant les conflits d'intérêts avec un SCL (*Guide de discussion* ; *Document technique*). Il est nécessaire d'avoir des mesures pour retirer des membres ou pour obliger une rotation afin d'éviter que ce soit toujours la même élite qui prend les décisions en matière de régulation du discours public en ligne.

Deuxièmement, lorsqu'il s'agit de prendre des décisions par rapport à un droit fondamental comme la liberté d'expression, l'idée de permettre au Conseil de recours de tenir des audiences en huis clos sème quelques risques (*Rapport synthèse*). Un huis clos signifie que seules les personnes concernées et nécessaires à la tenue de l'audience pourront y participer. De plus, seul le verdict sera public puisque tous éléments discutés en huis clos doivent demeurer privés. D'un côté, le huis clos permet le respect de la vie privée, la protection de la réputation des entités concernées et la protection de l'identité d'une personne. Cependant, d'un autre côté, le huis clos ne permet pas de pouvoir vérifier que l'audience a été tenue dans le bon respect des règles ni de voir les discussions qui ont mené au verdict. Il faut alors faire entièrement confiance au Conseil de recours. Dans une société démocratique, surtout quand vient le temps de trancher sur quel contenu peut ou ne peut pas être visible sur Internet, la population devrait pouvoir avoir accès à l'entièreté des processus décisionnels (*Rapport synthèse* ; OpenMedia, 2021, p. 14). Comme alternative au huis clos, il pourrait être proposé de plutôt rendre certains éléments non accessibles au public. Le Conseil de recours doit déjà en tout temps protéger l'identité du plaignant (*Document technique*, art. 83).

Enfin, l'Approche prévoit donner un pouvoir d'accès à l'information énorme au commissaire à la sécurité numérique. Ce dernier peut mener des inspections auprès des SCL à tout moment et « […] à son entière discrétion […] » (*Document technique*, art. 88). Les inspecteurs n'auront plus besoin de mandat pour entrer dans un lieu autre qu'une maison d'habitation afin d'effectuer une inspection et les personnes responsables de ce lieu devront fournir une assistance nécessaire pour répondre aux besoins des inspecteurs (*Document technique*, art. 89). En retirant la nécessité d'obtenir un

mandat pour recueillir des renseignements privés ou confidentiels, le gouvernement du Canada ouvre la porte à divers abus qui mettrait à risque le droit à la vie privée (*Rapport synthèse* ; Geist, 2021, p. 9).

En bref, bien qu'il soit nécessaire d'avoir des organismes de réglementation indépendants pour assurer l'application de la loi et l'évolution de cette dernière, il serait important que le gouvernement du Canada mette en place plus de mécanismes de redevabilité et s'assure de ne pas donner des pouvoirs non nécessaires qui pourraient nuire au droit à la vie privée des Canadiens et Canadiennes.

4.4.2 Équité

Au niveau du critère d'équité des nouveaux organismes de réglementation, l'enjeu majeur soulevé est le manque potentiel de diversité, élément qui semble revenir fréquemment dans l'Approche. Bien que le gouvernement du Canada mentionne nombreuses fois dans le *Guide de discussion* et le *Document technique* qu'il est primordial que les organismes représentent la diversité culturelle, de genre et d'expertise présente au Canada, il est difficile d'imaginer l'atteinte de cet objectif dû au fait que les organismes prévoient très peu de membres (*Rapport synthèse*). Le Conseil de recours est composé de trois à cinq membres et le comité d'experts, de son côté, prévoit un maximum de sept membres. Peu importe le nombre d'efforts investis pour obtenir des groupes diversifiés, trop peu d'intérêts, de cultures, d'expertises et de réalités socioéconomiques pourront être représentés au sein de 12 membres maximum (*Rapport synthèse*).

Le Canada est reconnu, entre autres, pour sa diversité culturelle. Chaque province et chaque région a ses propres différences culturelles qui rendent souvent l'obtention d'un consensus difficile. Cela dit, l'Approche prévoit mettre en place un cadre réglementaire universel, ce qui pourrait amener des enjeux au niveau de la diversité. Tout comme au sein des équipes des SCL, il y a de fortes chances que les intérêts de la majorité culturelle dominent et effacent ceux des communautés marginalisées (*Rapport synthèse*). Certaines solutions ont déjà été proposées dans les dernières années pour tenter de trouver des solutions plus inclusives à la modération de contenu. Parmi celles-ci, les Conseils de médias sociaux (*social media councils*) en sont un exemple (Center for International Governance Innovation, 2019, p. 99). Ces conseils adoptent une approche multipartite

en rassemblant, notamment, des représentants des plateformes, des groupes de la société civile avec des intérêts politiques divers et des représentants de groupes marginalisés (Article 19, 2021b ; Center for International Governance Innovation, 2019, p. 99). De plus, ceux-ci peuvent être régionaux, nationaux ou internationaux (Article 19, 2021b, p. 7). Il serait intéressant d'imaginer avoir un Conseil de recours national qui élabore les règles à suivre pour les différentes divisions régionales. Chaque région ou chaque province pourrait avoir sa propre division du Conseil de recours afin que les personnes qui en font partie soient conscientes et connaisseurs du contexte culturel spécifique de cette communauté (Article 19, 2021b, p. 7).

4.4.3 Faisabilité

Au niveau de la faisabilité, le Conseil de recours semble compter trop peu de membres pour pouvoir efficacement répondre à son mandat et aux besoins de la population canadienne (*Rapport synthèse* ; Geist, 2021, p. 10). À titre de rappel, le Conseil de recours en matière de sécurité numérique servirait comme une cour d'appel indépendante pour les citoyens suite aux décisions prises par les services de communication en ligne en matière de modération de contenu (*Guide de discussion* ; *Document technique*). Il serait composé de trois à cinq membres provenant d'expertises diversifiées et représentant la diversité culturelle du Canada (*Document technique, art. 46*). Puisqu'il n'existe aucun organisme de ce genre au Canada présentement, il est difficile d'estimer le nombre de plaintes ou d'appels qui seront soumis au Conseil de recours. Cependant, il est possible d'imaginer que cet organisme recevra probablement des milliers de cas annuellement. Cette tâche est colossale et quasi impossible à réaliser pour une équipe de trois à cinq membres (*Rapport synthèse* ; Geist, 2021, p. 10). Cela dit, au niveau de la faisabilité de cette mesure, il est possible d'imaginer qu'il deviendra tout simplement un nouveau processus bureaucratique lourd et inaccessible à la population canadienne (*Rapport synthèse*).

Faisons une comparaison avec le Conseil de surveillance de Facebook qui a été mis en place pour agir à titre de conseil indépendant afin de réviser certains cas de modération de contenu de Facebook et Instagram. Les usagers peuvent soumettre une demande d'appel une fois qu'ils ont utilisé tous les recours possibles au sein de Facebook (Conseil de surveillance, s.d.). Le Conseil de surveillance compte présentement 23 membres de nationalités et expertises diversifiées et souhaite atteindre un maximum de 40 membres. D'abord, le Conseil de surveillance ne revoit pas chaque

cas soumis. Il choisit les cas en fonction de leurs importances et si ces derniers peuvent établir un précédent pour la modification des standards de communauté de Facebook et Instagram (Conseil de surveillance, s.d.). Ensuite, quelques membres du Conseil sont choisis pour examiner le cas sélectionné (Conseil de surveillance, s.d.). En d'autres mots, les membres ne sont jamais tous assignés au même cas permettant d'en traiter plusieurs à la fois. Enfin, le Conseil de surveillance a jusqu'à 90 jours pour rendre une décision (Conseil de surveillance, s.d.).

Revenons au Conseil de recours en matière de sécurité numérique au Canada. Selon le *Guide de discussion* et le *Document technique*, le Conseil de recours doit traiter, tous ensemble, l'entièreté des plaintes soumises. Notons qu'il n'y a pas de mention de délai pour rendre les décisions. Premièrement, le peu de membres rend le processus laborieux, comme déjà mentionné auparavant. Deuxièmement, en comparant avec le Conseil de surveillance de Facebook qui prend jusqu'à 90 jours pour rendre une décision, il est possible d'imaginer des délais semblables pour les plaintes soumises au Conseil de recours. Les plaintes soumises seront probablement très rarement des cas de contenus clairement illégaux. Puisqu'elles seront des cas de contenus dans des zones grises, les membres du Conseil de recours devront analyser le contexte local et les intentions pour ne pas prendre des décisions à la hâte puisque ceci pourrait enfreindre la liberté d'expression ou autres droits des Canadiens et Canadiennes (Caplan, 2018, p. 13 ; Roberts, 2016, p. 2). Cela dit, le processus de traitement de plaintes sera certainement très long et certaines plaintes pourraient devenir désuètes avant même d'être traitées.

Le Conseil de recours pourrait fonctionner davantage comme le Conseil de surveillance de Facebook. Je tiens à souligner qu'il ne s'agit pas de dire que le modèle utilisé par Facebook est parfait, mais plutôt de reconnaître l'innovation de cette approche. Cela dit, le Conseil de recours aurait pour mandat de réviser quelques plaintes soumises qui permettront d'établir des précédents et de clarifier les limites de ce qui est considéré comme préjudiciable ou non. Ces décisions et recommandations, accessibles publiquement, devraient être prises en compte par les différentes entités réglementées pour l'amélioration de leurs standards de communauté.

Tableau 12 Points saillants de l'analyse des nouveaux organismes de réglementation

Risques	Équité	Faisabilité
Risque de créer un déficit démocratique et manque de redevabilité. Risque de voir des abus des nouveaux pouvoirs dus au manque de mécanismes de surveillance ou de redevabilité.	Potentiel manque de diversité causé par le peu de membres composant les nouveaux organismes de réglementation.	Les organismes de réglementation ne comptent pas suffisamment de personnes pour pouvoir bien répondre à leurs mandats.

5 SIMULATION D'UN MÉMOIRE DE CONSULTATION PUBLIQUE

Préambule : *Dans ce chapitre, je propose un commentaire en réponse à la consultation publique qui a eu lieu du 29 juillet au 25 septembre 2021 concernant l'Approche. Malgré le fait que cette consultation soit déjà terminée au moment d'écrire ce travail, je souhaite faire l'exercice tout de même afin de synthétiser mon analyse, rassembler l'information et proposer des recommandations pour l'amélioration de l'Approche.*

~~~~~~~~~~~~~~

EN RÉPONSE À L'APPROCHE PROPOSÉE PAR LE GOUVERNEMENT DU CANADA EN 2021 POUR LUTTER CONTRE LE CONTENU PRÉJUDICIABLE

Cher ministère du Patrimoine canadien,

J'écris en réponse à la consultation publique sur l'approche proposée par le gouvernement du Canada pour lutter contre le contenu préjudiciable. Certains éléments présentés dans le *Guide de discussion* et le *Document technique* semblaient être prometteurs au niveau de l'idée, comme offrir un conseil de recours indépendant pour les Canadiens et Canadiennes qui souhaitent faire appel aux décisions de modération de contenu prises par les SCL. Bien que l'initiative du gouvernement et les motivations derrière l'approche proposée pour minimiser la propagation de contenus préjudiciables en ligne soient louables, plusieurs problématiques sont à prendre en considération avant de déployer un tel projet de loi.

Ce cadre réglementaire et législatif aura un impact sur les environnements numériques et sur la vie sociale et politique des Canadiens et Canadiennes en définissant ce qui est compris comme un préjudice, en invisibilisant certains préjudices, en exacerbant des problèmes d'équité préexistants, en forçant des prises de décisions à la hâte et en ayant un organisme de réglementation autoritaire qui ne promet aucun changement concret.

1. Définir les préjudices

L'Approche définit cinq catégories de contenus préjudiciables alors qu'il n'y a aucun consensus sur ce qui constitue un préjudice et même sur ce qui devrait être régulé par un cadre réglementaire et législatif du gouvernement. En analysant les cadres réglementaires du Royaume-Uni, de
~~~~~~~~~~~~~~

l'Allemagne, de l'Australie et de la France, on constate que chaque pays cible différents types de contenus préjudiciables. Alors que le Royaume-Uni et la France considèrent la désinformation comme un contenu préjudiciable (République française, 2020 ; Department for Digital, Culture, Media & Sport, 2022), car celle-ci peut être la source d'incitation à la violence et de radicalisation, le gouvernement du Canada a décidé de l'exclure de l'Approche. Même les experts ne s'entendent pas sur la portée et les catégories de contenus préjudiciables qui devraient être pris en compte dans l'Approche (*Rapport synthèse*). Certains mentionnent que l'Approche cible un éventail trop diversifié de types de contenus pour un même cadre réglementaire considérant qu'ils ne sont pas évalués avec les mêmes critères sur le plan juridique (*Rapport synthèse* ; Geist, 2021, p. 2 ; Khoo *et al*., 2021, §17). D'autres considèrent que l'Approche ne devrait pas cibler de catégorie, car il est impossible de prévoir tous les types de préjudices qui peuvent être vécus en ligne (Rapport synthèse ; Ministère du Patrimoine canadien, 2022d).

Ensuite, parmi les catégories de contenu préjudiciable ciblées par le gouvernement du Canada, il est important de faire la distinction entre les discours haineux et les quatre autres catégories (le contenu terroriste, le contenu incitant à la violence, le partage non consensuel d'images intimes et le contenu d'exploitation sexuelle d'enfants en ligne). Contrairement aux discours haineux, les autres catégories de contenus ne dépendent pas nécessairement du contexte local ou culturel (Caplan, 2018, p. 13). Pour les discours haineux, il est complexe de savoir où tracer la ligne entre ce qui est illégal et ce qui est un discours protégé (par exemple, la satire ou la critique). Les modérateurs doivent être en mesure de distinguer les codes et référents culturels pour voir les nuances dans le discours (Roberts, 2016, p. 2 ; Caplan, 2018, p. 13). Le contenu aura toujours une fonction performative pouvant blesser quelqu'un juste par l'action qu'il accomplit (Butler, 2004, p. 108).

De plus, chaque décision prise concernant un discours de haine est une redéfinition des normes et une déclaration que ce langage est haineux (Gillespie, 2020, p. 3). Cela dit, les contenus préjudiciables sont en constant changement et évoluent pour se fondre dans le discours public. Il y a un besoin constant de réévaluer nos normes et valeurs de ce qui est considéré comme préjudiciable tout en prenant en compte le contexte et les référents culturels, car ceux-ci peuvent grandement influencer l'expérience que quelqu'un aura face à un type de discours. En

conséquence, avoir des définitions fixes de ce qui est considéré comme préjudiciable peut nuire à l'évolution de ces dernières qui agissent dans des environnements constamment en changement.

Recommandation : *Autant qu'il soit nécessaire d'avoir des définitions claires des contenus préjudiciables pour guider les SCL et les usagers, il est nécessaire de prévoir des possibilités de faire évoluer et d'adapter les catégories de contenus ciblées par le cadre réglementaire et législatif afin de refléter les différents types de préjudices vécus par la population canadienne. Dans cette même idée, l'utilisation d'une approche basée sur le risque peut être explorée comme alternative (*Ministère du Patrimoine canadien, 2022e*). Le gouvernement du Canada obligerait les SCL à évaluer les risques de circulation de contenus préjudiciables sur leurs plateformes. Les SCL devront mettre en place des mesures adéquates, proportionnelles et spécifiques aux risques qui les concernent. Ce processus serait, bien entendu, surveillé par le cadre réglementaire pour assurer que les SCL répondent à leur « devoir de diligence » (*Ministère du Patrimoine canadien, 2022e*). Cette alternative ne définirait pas des catégories de contenus préjudiciables ex ante permettant aux SCL de s'adapter aux différentes formes de préjudices (*Ministère du Patrimoine canadien, 2022e*).*

2. Invisibiliser les préjudices

L'Approche mènerait à l'invisibilisation de certains préjudices. Premièrement, l'acte de catégoriser les préjudices signifie que certains types de préjudices ne sont pas reconnus sous la loi. En ce sens, certains préjudices seront invisibilisés et minimisés par le fait même de ne pas être inclus dans le cadre réglementaire et législatif.

Deuxièmement, en ignorant les risques de la propagation de contenus préjudiciables sur les services de communication privée ou cryptée, le gouvernement du Canada incite la migration de ce contenu sur des plateformes non réglementées (Dugal et Lozach, 2021). Après l'assaut du Capitole aux États-Unis le 6 janvier 2021 et le mouvement de *deplatforming,* plusieurs groupes extrémistes ont migré vers des plateformes de communication privée et cryptée qui modèrent très peu le contenu (Dugal et Lozach, 2021). Par exemple, plusieurs usagers ont migré vers la plateforme *Telegram*. La taille des groupes de discussion « privée » sur *Telegram* peut aller jusqu'à 200 000 personnes

(Dugal et Lozach, 2021). Est-ce toujours considéré comme des communications privées ? Le gouvernement du Canada devra déterminer où tracer la ligne entre privé et public.

En excluant entièrement les services de communication privée, le gouvernement du Canada protège le droit à la vie privée, mais prend un grand risque de tout simplement cacher le contenu préjudiciable. Au lieu de réellement limiter le partage de ce contenu et de réduire le volume de ce genre de contenu en ligne, l'approche proposée par le gouvernement du Canada pourrait tout simplement le faire migrer vers des SCL non réglementés et en conséquence, invisibiliser certains préjudices.

Recommandation : *Le gouvernement du Canada devrait se pencher sur les différentes mesures possibles pour minimiser les risques de propagation du contenu préjudiciable sur les services de communication privée et cryptée, sans pour autant faire la surveillance des messages privés de la population canadienne. À titre d'exemples, ces mesures ressemblent à obliger la mise en place de limites sur la taille des groupes de discussion, de limites sur le nombre de fois que les contenus peuvent être transférés, de mécanismes de signalement par les usagers et d'outils de vérification de l'information reçue (Andrey et al., 2021).*

3. Exacerber les biais préexistants impactant disproportionnellement les communautés marginalisées

Dans le *Guide de discussion* et le *Document technique*, le gouvernement du Canada recommande fortement l'utilisation d'outils de modération de contenu automatisés pour assurer le respect des nouvelles règles et obligations légales. D'abord, l'exercice de définir les standards de communauté et les règles par rapport à la modération de contenu est un acte subjectif qui reflète les valeurs, les normes et les biais des décideurs au sein des SCL. Ceux-ci sont ensuite intégrés dans le code des outils d'automation, retirant leur « neutralité » (Gerrard et Thornham, 2020, p. 1271).

De plus, les outils d'automation principalement utilisés par les services de communication en ligne relèvent de l'apprentissage automatique (*machine learning*). L'apprentissage automatique tente simplement de trouver des correspondances entre le contenu publié et des contenus qui ont déjà été catégorisés comme « non désirés » (Gerrard and Thornham, 2020, p. 1269; Gillespie, 2020). Cela

dit, ces outils d'automation figent dans le temps ce qui est acceptable ou non (Binns *et al*., 2017, p. 7). Puisque l'élaboration des standards de communauté et des règles de modération de contenu est subjectif, les bases de données et les codes des outils sont souvent empreints de biais historiques qui ne cessent d'être recréés à travers le processus de modération de contenu algorithmique. Ce sont des biais qui impactent disproportionnellement les communautés marginalisées parce qu'elles ne sont pas adéquatement représentées au sein des équipes.

Au-delà des outils d'automation, les mécanismes de signalement sur les SCL sont souvent l'objet de ludification (Crawford et Gillespie, 2016, p. 420). À travers ces jeux de signalement, certains groupes marginalisés peuvent être ciblés par des actions coordonnées d'usagers qui souhaitent limiter leur participation aux discours publics, en rendant leur contenu invisible et par conséquent, retirant la légitimité du contenu (Crawford et Gillespie, 2016, p. 421). En bref, il s'agit d'un risque de voir des communautés marginalisées être réduites au silence par une majorité culturelle.

Recommandation : *Le gouvernement du Canada devrait sensibiliser les SCL face aux impacts que peuvent avoir les outils d'automation sur les communautés marginalisées. L'utilisation de modérateurs humains et l'importance d'avoir des équipes techniques qui représentent la diversité culturelle, ethnique et de genre de la population canadienne devraient être priorisées par l'Approche. De plus, un travail d'éducation de la population au sujet de la modération de contenu communautaire devrait être entrepris pour habiliter les Canadiens et Canadiennes à participer activement au façonnement de leurs environnements numériques.*

4. Prendre des décisions complexes à la hâte dues au délai de 24 h

Le gouvernement du Canada mentionne dans le *Guide de discussion* et le *Document technique* que les contenus préjudiciables devront être rendus inaccessibles au Canada dans un délai de 24 heures suivant le signalement initial. Ce délai, assez court, obligera les services de communication en ligne à prendre des décisions de modération de contenu très rapidement. Pour ce qui est du contenu clairement illégal, je pense qu'il n'est pas irraisonnable de s'attendre à ce que ce délai soit respecté surtout considérant que ce genre de contenu est souvent bloqué *ex ante*. Je me soucis plutôt du contenu qui se retrouveront dans des zones grises et qui nécessiteraient plus de temps de réflexions sur le contexte local et l'intention derrière le contenu. Afin d'éviter les sanctions, les SCL

pencheront vers la prudence et retireront probablement du contenu qui n'aurait peut-être pas été considéré comme préjudiciable si l'équipe de modération avait eu plus de temps pour en faire l'évaluation.

Ce délai de 24 h pourrait aussi potentiellement nuire à la détection de nouveau cas de contenu préjudiciable. Comme mentionné précédemment, ce genre de contenu a tendance à prendre de nouvelles formes afin de contourner les mécanismes de modération de contenu. Par exemple, les usagers qui souhaitent partager du contenu pro trouble alimentaire ont commencé à utiliser des symboles pour signaler leur identification à cette communauté (Gerrard, 2018). Puisqu'il s'agit de symboles qui font sens uniquement aux membres de cette communauté, les outils d'automation ou les modérateurs humains ne les détectent pas. Cela dit, afin de pouvoir comprendre et détecter ces mécanismes de contournement de la modération de contenu, les modérateurs doivent avoir suffisamment de temps pour se pencher sur l'évaluation du contenu en question. Un risque majeur d'imposer un délai de 24 heures aux SCL est de manquer ces évolutions du contenu préjudiciable parce que les modérateurs sont toujours pressés de prendre des décisions.

Recommandation : *Pour certains, l'idéal serait de ne pas imposer une limite de temps pour rendre le contenu inaccessible au Canada puisque ceci pourrait être irréalisable pour certains petits SCL et pourrait mener à une suppression excessive du contenu impactant la liberté d'expression des usagers (Rapport synthèse ;* Baghdasaryan et Gullo, 2021 ; Article 19a, 2021). *De mon côté, j'invite le gouvernement du Canada à être flexible dans les délais alloués. Tout comme le NetzDG en Allemagne, les délais pourraient varier selon la nature du contenu préjudiciable. De plus, il pourrait être requis des services de communication en ligne d'intégrer dans les rapports de transparence leur délai de réponse face aux divers types de contenus préjudiciables en ligne afin d'évaluer si les SCL mettent suffisamment d'efforts dans la lutte pour freiner la propagation et la visibilité des contenus préjudiciables.*

5. Un Conseil de recours autoritaire qui n'apporte aucun changement concret s'il réussit à relever son mandat

Le Conseil de recours en matière numérique du Canada deviendra un organisme indépendant, autoritaire, non redevable à la population et aura de la difficulté à réaliser son mandat. D'abord, le

Document technique mentionne très peu de détails sur le processus de sélection des membres du Conseil de recours. L'unique élément mentionné est le fait qu'ils seront nommés par le ministère du Patrimoine canadien. Est-ce qu'il y a une rotation prévue ou sont-ils nommés à vie ? Est-ce qu'il y a des mécanismes prévus pour les conflits d'intérêts ? Est-ce qu'il y a une façon de retirer un membre qui ne représenterait pas adéquatement les valeurs, normes et intérêts de la population canadienne ? Sont-ils redevables à la population canadienne considérant qu'ils prennent des décisions qui impactent directement la liberté d'expression ? La façon dont le Conseil de recours est présenté soulève des risques face au processus démocratique et donne des pouvoirs énormes, sans garde-fous afin de s'assurer que la liberté d'expression des Canadiens et Canadiennes soit respectée.

Ensuite, le Conseil de recours prévoit être composé de trois à cinq membres pour répondre aux plaintes provenant de partout au Canada. Cette tâche est colossale et quasi impossible pour si peu de membres. Considérant que le Conseil de recours traitera toutes les plaintes reçues, il est difficile d'envisager qu'il réussira à adéquatement répondre à son mandat.

Enfin, le Conseil de recours n'a pas de réels pouvoirs pour produire du changement au sein des SCL. En bref, les décisions prises concernent uniquement l'obligation de retirer du contenu préjudiciable. Cependant, ce processus n'amènera aucun changement concret puisqu'il s'agit uniquement de corriger des décisions prises par les SCL. Pour voir de réels changements, le Conseil de recours pourrait émettre des recommandations aux SCL afin qu'ils améliorent leurs processus de modération de contenu ou qu'ils modifient leurs termes et conditions d'utilisation afin d'éviter de toujours revoir les mêmes cas revenir sans cesse.

Recommandation : *Plusieurs recommandations peuvent être émises pour améliorer la structure du Conseil de recours en matière numérique du Canada.*

a. *Le Conseil de recours pourrait avoir des divisions régionales. Ceci permettrait de diviser le mandat entre plusieurs entités, de plus facilement considérer les différences culturelles importantes d'une région à l'autre et de représenter plus de diversité au sein de l'organisme de réglementation.*

b. *Le Conseil de recours pourrait avoir une approche multipartite pour davantage inclure les différents acteurs impliqués et impactés par la modération de contenu. Par exemple, au sein du conseil, on pourrait retrouver des membres du secteur privé, d'organismes non gouvernementaux et des représentants de la société civile.*
c. *Le Conseil de recours pourrait avoir un mandat et des pouvoirs similaires au Conseil de surveillance de Facebook (Oversight board). Le Conseil de surveillance est indépendant et a été mis en place pour réviser les décisions de modération de contenu prises par la compagnie (Conseil de surveillance, s/d). Il ne traite aucunement toutes les plaintes reçues. Au contraire, une sélection de quelques cas importants sont sélectionnés afin d'établir des précédents au niveau des décisions. Le Conseil de surveillance peut prendre jusqu'à 90 jours pour rendre une décision et celle-ci est contraignante signifiant que Facebook doit appliquer le jugement rendu. Enfin, chaque décision est accompagnée de recommandations pour l'amélioration des processus de modération de contenu et des standards de communauté. Facebook n'a pas d'obligation d'appliquer les recommandations, mais doit y répondre et justifier leur décision. Ce modèle, quoiqu'imparfait, pourrait être une source d'inspiration pour l'amélioration du Conseil de recours en matière numérique du Canada.*

En conclusion, cette réponse à la consultation publique met en lumière quelques risques et enjeux en lien avec l'Approche. Il n'est pas question de complètement rejeter ce qui a été proposé par le gouvernement du Canada, mais il est nécessaire de prendre en considération les problématiques soulevées pour améliorer l'Approche. Malgré tous les avancements technologiques, il n'existe pas de solution miracle pour modérer le contenu préjudiciable en ligne. Il est important d'avoir des discussions publiques sur ces enjeux et d'inclure davantage les usagers dans ces prises de décisions puisque leurs droits fondamentaux seront directement impactés.

6 CONCLUSION

En conclusion, ce travail dirigé m'a permis de mettre en lumière plusieurs enjeux de la modération de contenu sur Internet comme les biais empreints dans les outils d'automation, l'importance de prendre en compte les contextes locaux dans le processus décisionnel, les difficultés d'avoir une approche universelle et les différents modèles de modération de contenus utilisés par les médias sociaux. L'analyse de l'Approche démontre aussi la complexité d'élaborer un cadre réglementaire et législatif en matière de modération de contenu et l'importance d'avoir l'avis d'experts de divers domaines pour minimiser les impacts négatifs et maximiser l'efficacité de cette dernière. Au final, l'Approche est loin d'être une solution parfaite. Toutefois, elle permet d'ouvrir une discussion au Canada au sujet d'impliquer les instances gouvernementales dans cette mission de rendre l'Internet plus sécuritaire et inclusif.

De mon côté, effectuer une analyse de politique de l'Approche m'a permis d'explorer en détail les différents éléments qui doivent être pris en compte lorsqu'on parle de réguler les SCL. Le but de ce travail n'était pas de trouver la solution qui réglerait tous les problèmes, mais plutôt de proposer de nouvelles pistes de réflexion pour l'élaboration de nouveaux cadres réglementaires et législatifs comme j'ai fait dans la simulation du mémoire de consultation publique. À travers mon processus de recherche, j'ai réussi à obtenir, en faisant une demande d'accès à l'information, l'ensemble des réponses envoyées au ministère du Patrimoine canadien dans le cadre de la consultation publique. Si j'étais pour faire suite à ce travail dirigé, il saurait intéressant de faire une analyse du discours entourant l'Approche. Cette étude pourrait inclure l'analyse des mémoires de consultation publique pour avoir une idée de l'opinion des différents acteurs au Canada, l'analyse des conférences de presse données par le gouvernement et l'opinion publique des usagers concernant la régulation du contenu en ligne par le gouvernement du Canada.

Pour finir, ayant choisi une méthodologie peu utilisée dans le domaine des communications, j'ai ressenti un syndrome d'imposteur pendant la rédaction de ce travail dirigé. Cependant, après maintes discussions et réflexions, je crois fermement que ce genre d'analyse devrait être davantage effectué dans notre domaine. Pour tous types de régulation, il est important que divers experts se penchent sur les solutions proposées, car chaque domaine permet de soulever des risques et enjeux que d'autres n'auraient probablement pas remarqués. Au moment d'écrire la dernière phrase de

mon travail dirigé, je peux affirmer que mon analyse de l'approche proposée par le gouvernement du Canada pour lutter contre le contenu préjudiciable en ligne a sa place dans le domaine des communications et j'espère que d'autres étudiants et étudiantes se pencheront sur des questions de régulation au Canada.

7 BIBLIOGRAPHIE

Access Now. (2021, 24 septembre). Re: The Government of Canada's proposed approach to address harmful content online. Dans le cadre de la consultation publique du gouvernement du Canada. https://www.accessnow.org/cms/assets/uploads/2021/09/Access-Now-Canada-Online-Harms-Proposal-Comments-Final-09232021.pdf

Andrey, S., Rand, A., Masoodi, M. J., et Tran,S. (2022, mai). Messagerie privée, Préjudices publics. https://www.cybersecurepolicy.ca/privatemessaging

Article 19. (2021a, 26 juillet). UK: Draft Oline Safety Bill poses serious risk to free expression. *Article 19*. https://www.article19.org/resources/uk-draft-online-safety-bill-poses-serious-risk-to-free-expression/

Article 19. (2021b, 12 octobre). Social Media Councils : One piece in the puzzle of content moderation. *Article 19*. https://www.article19.org/wp-content/uploads/2021/10/A19-SMC.pdf

Badouard, R. (2021). Modérer la parole sur les réseaux sociaux : politiques des plateformes et régulation des contenus. *Réseaux*. 1(225), 87-120. 10.3917/res.225.0087

Baghdasaryan, M. et Gullo, K. (2021, 23 novembre). UN Human Rights Committee Criticizes Germany's NetzDG for Letting Social Media Platforms Police Online Speech. *Electronic Frontier Foundation*. https://www.eff.org/deeplinks/2021/11/un-human-rights-committee-criticizes-germanys-netzdg-letting-social-media

Banchik, A. V. (2021). Disappearing acts: Content moderation and emergent practices to preserve at-risk human rights–related content. *New Media & Society*, *23*(6), 1527-1544.

Bozio, A. (2018). Les méthodes d'évaluation des politiques publiques. *Idées économiques et sociales*. 3(193), 28-33. 10.3917/idee.193.0028

Brown, A. (2018). What is so special about online (as compared to offline) hate speech?. *Ethnicities*. 18(3), 297-326. 10.1177/1468796817709846

Butler, J. (2004). Le pouvoir des mots : Discours de haine et politique du performatif. Éditions Amsterdam.

Canadian Citizens' Assembly on Democratic Expression. (2021). Canadian Citizens' Assembly on Democratic Expression: Recommendations to strengthen Canada's response to new digital technologies and reduce the harm caused by their misuse. Ottawa, Public Policy Forum. https://ppforum.ca/wp-content/uploads/2021/01/CanadianCitizens%E2%80%99AssemblyOnDemocraticExpression-PPF-JAN2021-EN.pdf

Caplan, R. (2018, novembre). Content or context moderation? Artisanal, Community-Reliant, and Industrial Approaches. *Data & Society.* https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf

Centre canadien de protection de l'enfance. (2016, janvier). Les images d'abus pédosexuels sur Internet : Une analyse de Cyberaide.ca. https://www.protectchildren.ca/pdfs/CTIP_CSAResearchReport_2016_fr.pdf

Centre canadien de protection de l'enfance. (2021, 8 juin). Projet Arachnid : l'accessibilité des images d'abus pédosexuels sur Internet (Document de synthèse). https://protectchildren.ca/pdfs/C3P_ProjectArachnidReport_Summary_fr.pdf

Center for Democracy & Technology. (2017, juillet). Overview of the NetzDG Network Enforcement Law. https://cdt.org/insights/overview-of-the-netzdg-network-enforcement-law/

Centre for International Governance Innovation. (2019). Models for Platform Governance: A CIGI Essay Series. https://www.cigionline.org/models-platform-governance/

Commission européenne. (2016). Code of conduct on countering illegal hate speech online. https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

Conseil de surveillance. (s/d). Comment fonctionne le processus d'appel. https://oversightboard.com/appeals-process/

Cook, C. L., Patel, A., et Wohn, D. Y. (2021). Commercial versus volunteer: Comparing user perceptions of toxicity and transparency in content moderation across social media platforms. *Frontiers in Human Dynamics*, 3, 3.

Crawford, K., et Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410-428.

Department for Digital, Culture, Media & Sport. (2022, 19 avril). Policy paper: Online Safety Bill: factsheet. Gov.uk. https://www.gov.uk/government/publications/online-safety-bill-supporting-documents/online-safety-bill-factsheet

Department for Digital, Culture, Media & Sport et Dorries, The Rt Hon N. (2022, 17 mars). Press release: World-first online safety laws introduced in Parliament. https://www.gov.uk/government/news/world-first-online-safety-laws-introduced-in-parliament#:~:text=The%20Online%20Safety%20Bill%20marks,while%20protecting%20freedom%20of%20speech

DeNardis, L., et Hackl, A. M. (2015). Internet governance by social media platforms. *Telecommunications Policy*, *39*(9), 761-770.

Dugal, M. et Lozach, F. (2021, 2 novembre). Pourquoi certaines applications sont-elles appréciées des conspirationnistes? *Moteur de recherche [Balado]. Radio-Canada.* https://ici.radio-canada.ca/ohdio/premiere/emissions/moteur-de-recherche/segments/chronique/377438/telegram-qanon-trump

eSafety Commissioner. (2022, janvier). Online Safety Act 2021: Fact sheet. *Australian Gouvernment*. https://www.esafety.gov.au/sites/default/files/2021-07/Online%20Safety%20Act%20-%20Fact%20sheet.pdf

Fortin, M. F., & Gagnon, J. (2016). Fondements et étapes du processus de recherche: méthodes quantitatives et qualitatives. *Chenelière éducation*.

Franzke, A.S., Bechmann, A., Zimmer, M., Ess, C. et the Association of Internet Researchers. (2020). Internet Research: Ethical Guidelines 3.0. https://aoir.org/reports/ethics3.pdf

Geist, M. (2021, septembre). Government of Canada Consultation on the Proposed Approach to Address Harmful Content Online. Dans le cadre de la consultation publique du gouvernement du Canada. https://www.michaelgeist.ca/wp-content/uploads/2021/09/Geistonlineharmssubmission.pdf

Gellert, R. (2018). Understanding the notion of risk in the General Data Protection Regulation. *Computer Law & Security Review*, 34(2), 279-288.

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, *20*(12), 4492-4511.

Gerrard, Y. et Thornham, H. (2020). Content moderation: Social media's sexist assemblages. *New Media & Society*, 22(7), 1266-1286.

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2). http://dx.doi.org/10.1177/2053951720943234

Godbout, M. (2022, 3 février). Réforme de la Loi sur la radiodiffusion : prise deux. *Radio-Canada*. https://ici.radio-canada.ca/nouvelle/1859347/crtc-reforme-loi-radio-tele-geants-web-contenu-canadien

Goosz, Y. (2021, 11 janvier). Haine en ligne : une deuxième vie pour la loi Avia?. *France Inter*. https://www.franceinter.fr/politique/haine-en-ligne-une-deuxieme-vie-pour-la-loi-avia

Gorwa, R., Binns, R., et Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1). 10.1177/2053951719897945.

Gouvernement du Canada. (2021, juillet). L'exploitation sexuelle des enfants en ligne : un fléau en hausse au Canada [*infographie*]. https://www.canada.ca/content/dam/ps-sp/documents/campaigns/online-child-sexual-exploitation/OCSE_Infographic_fr.pdf

Grădinaru, D. (2018). The Principle of Legality. *Research Association for Interdisciplinary studies*. 10.5281/zenodo.1572191

Groupe de travail du comité de coordination des hauts fonctionnaires sur le cybercrime. (2013, juin). Rapport aux ministres fédéraux/provinciaux/territoriaux responsables de la Justice et de la Sécurité publique : Cyberintimidation et distribution non consensuelle d'images intimes. *Ministère de la Justice du Canada*. https://www.justice.gc.ca/fra/pr-rp/autre-other/cdncii-cndii/pdf/cdncii-cndii-fra.pdf

Heldt, A. (2019). Reading between the lines and the numbers: an analysis of the first NetzDG reports. *Internet Policy Review*, 8(2). 10.14763/2019.2.1398

Housefather, A. (2019). Agir pour mettre fin à la haine en ligne: Rapport du Comité permanent de la justice et des droits de la personne. *Chambre des communes*. 42e législature, 1re session. https://www.noscommunes.ca/Content/Committee/421/JUST/Reports/RP10581008/justrp29/justrp29-f.pdf

Howlett, M. et Lindquist, E. (2004). Policy Analysis and Governance: Analytical and Policy Styles in Canada. *Journal of Comparative Policy Analysis: Research and Practice*. 6(3), 225-249. 10.1080/1387698042000305194.

Internet Society Canada Chapter. (2021, 25 septembre). Submission to the Department of Canadian Heritage: Consultation on Internet Harms. Dans le cadre de la consultation publique du gouvernement du Canada. https://internetsociety.ca/wp-content/uploads/2021/09/ISCC-Response-Online-Harms-Final-21-9-21-1.pdf

Jhaver, S., Ghoshal, S., Bruckman, A., et Gilbert, E. (2018). Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *25*(2), 1-33.

Johnson, M., Mishna, F., Okumu, M., et Daciuk, J. (2018). Le partage non consensuel de sextos : Comportements et attitudes des jeunes Canadiens. *HabiloMédias*. http://habilomedias.ca/recherche-et-politique

Khoo, C., Gill, L., Parsons, C. et Citizen Lab. (2021, 25 septembre). Submission of the Citizen Lab to the Federal Government's Proposed Approach to Address Harmful Content Online ("Online Harms Consultation"). Dans le cadre de la consultation publique du gouvernement du Canada. https://citizenlab.ca/wp-content/uploads/2021/09/Citizen-Lab_Online-Harms-Consultation_25-Sept-2021_Letterhead_FINAL.pdf

Kieffer, A. et Laurent, M. (animateurs). (2020, 18 juin). *La loi Avia contre la haine en ligne censurée par le Conseil constitutionnel* [balado audio]. France culture. https://www.franceculture.fr/emissions/journal-de-18h/journal-de-18h-emission-du-jeudi-18-juin-2020

L'avenir nous appartient. (s.d.). Croisade contre la haine en ligne, le havre de paix inusité de Joanne Liu. *Télé-Québec*. https://lavenirnousappartient.telequebec.tv/emissions/333042/croisade-contre-la-haine-en-ligne-le-havre-de-paix-inusite-de-joanne-liu/63631/croisade-contre-la-haine-en-ligne-le-havre-de-paix-inusite-de-joanne-liu

La Presse canadienne. (2021a, 23 juin). Le projet de loi C-36 déposé aux Communes. *La Presse*. https://www.lapresse.ca/actualites/politique/2021-06-23/discours-haineux-sur-l-internet/le-projet-de-loi-c-36-depose-aux-communes.php

La Presse canadienne. (2021b, 27 novembre). Projet de loi C-10 : comment réglementer les services de diffusion en continu? *Radio-Canada*. https://ici.radio-canada.ca/nouvelle/1843223/reglementer-services-diffusion-continu-crtc-loi-federale

Milovanovitch, M. (2018). Guide de l'analyse des politiques. *Fondation européenne pour la formation*. 10.2816/01736. https://www.etf.europa.eu/sites/default/files/2018-07/Guide%20to%20policy%20analysis_FR.pdf

Ministère du Patrimoine canadien. (2021a). Guide de discussion. *L'engagement du gouvernement en faveur de la sécurité en ligne*. https://www.canada.ca/fr/patrimoine-canadien/campagnes/contenu-prejudiciable-en-ligne/guide-discussion.html

Ministère du Patrimoine canadien. (2021b). Document technique. *L'engagement du gouvernement en faveur de la sécurité en ligne*. https://www.canada.ca/fr/patrimoine-canadien/campagnes/contenu-prejudiciable-en-ligne/document-travail-technique.html

Ministère du Patrimoine canadien. (2021c). Créer un environnement numérique sûr, inclusif et transparent [*communiqué de presse*]. https://www.canada.ca/fr/patrimoine-canadien/nouvelles/2021/07/creer-un-environnement-numerique-sur-inclusif-et-transparent.html

Ministère du Patrimoine canadien. (2022a). Ce que nous avons entendu : Approche proposée du gouvernement pour s'attaquer au contenu préjudiciable en ligne. https://www.canada.ca/fr/patrimoine-canadien/campagnes/contenu-prejudiciable-en-ligne/ce-que-nous-avons-entendu.html

Ministère du Patrimoine canadien. (2022b). L'engagement du gouvernement en faveur de la sécurité en ligne. Gouvernement du Cananda. https://www.canada.ca/fr/patrimoine-canadien/campagnes/contenu-prejudiciable-en-ligne.html

Ministère du Patrimoine canadien. (2022c). Résumé de la première séance : les entités assujetties à la réglementation. *Gouvernement du Canada*. https://www.canada.ca/fr/patrimoine-canadien/campagnes/contenu-prejudiciable-en-ligne/premiere-seance-resume.html

Ministère du Patrimoine canadien. (2022d). Résumé de la deuxième session : Types de contenu à réglementer. *Gouvernement du Canada*. https://www.canada.ca/fr/patrimoine-canadien/campagnes/contenu-prejudiciable-en-ligne/premiere-seance-resume.html

Ministère du Patrimoine canadien. (2022e). Résumé de la cinquième séance : approche basée sur le risque. *Gouvernement du Canada*. https://www.canada.ca/fr/patrimoine-canadien/campagnes/contenu-prejudiciable-en-ligne/resume-cinquieme-seance.html

Morestin, F. et Centre de collaboration nationale sur les politiques publiques et la santé (CCNPP). (2012). Un cadre d'analyse de politique publique : guide pratique (1635). *Note documentaire : Pour des connaissances en matière de politiques publiques favorables à la santé*.

Mouketou, D. P. (2021). La lutte contre les contenus haineux sur les plateformes de médias sociaux: une analyse comparative d'approches de régulation [Rapport étudiant, École nationale d'administration publique (ENAP)], Espace ENAP. https://espace.enap.ca/id/eprint/250

Muller, P. (2000). L'analyse cognitive des politiques publiques : vers une sociologie politique de l'action publique. Revue française de science politique, 50(2), 189-207.

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11), 4366-4383.

OpenMedia. (2021, 25 septembre). Response to 'The Government's proposed approach to address harmful content online'. Dans le cadre de la consultation publique du gouvernement du Canada. https://openmedia.org/article/item/openmedias-response-to-the-harmful-content-consultation

Patton, C., Sawicki, D. et Clark, J. (2016). Basic Methods of Policy Analysis and Planning. *Taylor and Francis*. 3rd ed.

Rabeau, A. (s.d.). Final report: Regulatory Frameworks. Royal Commission on Worker's Compensation in British Columbia. http://www.qp.gov.bc.ca/rcwc/research/intersol-frameworks.pdf

Renaissance numérique. (2020, juin). Modération des contenus : Renouveler l'approche de la régulation. *Politiques, Institutions et Démocratie*. https://www.renaissancenumerique.org/system/attach_files/files/000/000/281/original/RenaissanceNumerique_Note_ModerationContenus.pdf?1613557546

République française. (2020, 25 juin). LOI n° 2020-766 du 24 juin 2020 visant à lutter contre les contenus haineux sur internet (1). *Journal officiel de la république française*. https://www.legifrance.gouv.fr/download/pdf?id=CP05NSqcPl5lPNu3MsP2PSu1fmt64dDetDQxhvJZNMc=

Riedl, M. J., Whipple, K. N. et Wallace, R. (2021). Antecedents of support for social media content moderation and platform regulation: the role of presumed effects on self and others. *Information, Communication & Society*. 10.1080/1369118X.2021.1874040

Roberts, S. T. (2016). Commercial content moderation: Digital laborers' dirty work. *Media Studies Publications*, 12. https://ir.lib.uwo.ca/commpub/12

Roberts, S. T. (2017). Social media's silent filter. *The Atlantic*, *8*.

Roberts, S. T. (2018, 5 mars). Digital detritus: 'Error' and the logic of opacity in social media content moderation. *First Monday*, 23(3). https://journals.uic.edu/ojs/index.php/fm/article/download/8283/6649

Sandelowski, M. (2000). Whatever Happened to Qualitative Description. *Research in Nursing & Health*, 23, 334-340.

Sécurité publique Canada. (2022). L'exploitation sexuelle des enfants sur Internet. https://www.securitepublique.gc.ca/cnt/cntrng-crm/chld-sxl-xplttn-ntrnt/index-fr.aspx

Shulock, N. (1999). The Paradox of PolicyAnalysis: If It Is Not Used Why Do We Produce So Much of It?. *Journal of Policy Analysis and Management*, 18(2), 226-244.

The Parliament of Australia (2021, 23 juillet). Online Safety Act 2021. https://www.legislation.gov.au/Details/C2021A00076

Tworek, H. et Leersen, P. (2019). An Analysis of Germany's NetzDG Law. *Transatlantic working group*.

Vie publique. (2020, 29 juin). Loi du 24 juin 2020 visant à lutter contre les contenus haineux sur internet. *Vie publique*. https://www.vie-publique.fr/loi/268070-loi-avia-lutte-contre-les-contenus-haineux-sur-internet

Vuong, K. (2022, 28 mars). Projet de loi C-261 : Loi modifiant le Code criminel, la Loi canadienne sur les droits de la personne et apportant des modifications connexes à une autre loi (propagande haineuse, crimes haineux et discours haineux). *Gouvernement du Canada*. https://www.parl.ca/Content/Bills/441/Private/C-261/C-261_1/C-261_1.PDF