La technologie, l'humain et le droit

Florence Guillaume (éd.)

droit.ius.tech

Stämpfli Editions

Information bibliographique de la Deutsche Nationalbibliothek
La Deutsche Nationalbibliothek a répertorié cette publication dans la Deutsche Nationalbibliografie; les données bibliographiques détaillées peuvent être consultées sur Internet à l'adresse http://dnb.d-nb.de.

Conception graphique de la couverture : deValence, Paris

www.staempfliverlag.com

Print ISBN 978-3-7272-4933-4

Dans notre librairie en ligne www.staempflishop.com,
la version suivante est également disponible :
E-Book ISBN 978-3-7272-4740-8

Table des matières

La modération de contenu par les réseaux sociaux

Les droits procéduraux des utilisateurs à la merci du pouvoir décisionnel des plateformes numériques

LEONEL CONSTANTINO FERREIRA
Assistant doctorant en droit international privé et droit des successions | LexTech Institute | Faculté de droit | Université de Neuchâtel

Table des matières

I. Introduction

En 2018, trois millions de publications ont été quotidiennement analysées par *Facebook* ; 300 000 ont été maintenues ou supprimées à tort par le réseau social, ce qui traduit un taux d'erreur de 10 %[1]. Il faut dire que la modération de contenu est un exercice complexe. Le terme « modérer » est défini comme l'action de diminuer l'intensité d'un phénomène et de réduire à sa juste mesure ce qui excessif[2]. Sur les réseaux sociaux, l'activité de modération vise à réduire, souvent par le biais d'algorithmes, l'intensité de la parole des utilisateurs. Mais comment déterminer ce qui est excessif ? Qui détient le pouvoir pour ce faire ? L'État et son appareil coercitif ou les réseaux sociaux et leurs conditions générales ? Doit-on interdire les contenus haineux, la nudité, la désinformation ou censurer l'utilisateur qui déclare, sur un ton sarcastique, qu'il va faire exploser un aéroport en raison de la fermeture inattendue de ce dernier[3] ?

La modération de contenu désigne les activités entreprises par les plateformes numériques destinées à détecter et à supprimer les contenus illicites ou indésirables publiés par des tiers (*i.e.* les utilisateurs)[4]. Les réseaux sociaux adoptent des mécanismes de modération dans le but de réguler la parole de leurs utilisateurs et ainsi assurer à ces derniers qu'ils ne soient pas confrontés à des contenus illicites ou choquants. Ces mécanismes ont une nature « hybride », dans le sens où ils combinent les outils technologiques et l'expertise humaine dans le but d'apprécier le caractère illicite d'une publication. Des algorithmes prédictifs sont notamment utilisés pour repérer et supprimer certains contenus.

Les législateurs nationaux (et supranationaux) ont fixé, dans leur droit positif, des régimes de responsabilité des plateformes numériques pour le contenu publié par leurs utilisateurs (responsabilité indirecte). Les régimes de responsabilité indirecte déterminent à quelles conditions une plateforme numérique est responsable pour ne pas avoir supprimé les contenus illicites publiés par ses

1 P. BARRETT, *Who Moderates the Social Media Giants ? A Call to End Outsourcing*, 2020, https://issuu.com/nyusterncenterforbusinessandhumanri/docs/nyu_content_moderation_report_final_version/1, consulté le 31.08.2022.

2 Dictionnaire Le Petit Robert, Paris 2022.

3 Cet exemple fait référence au *Twitter joke trial*, dans lequel un utilisateur a été condamné par les tribunaux britanniques pour « propos menaçants », en raison de la publication du *tweet* suivant : « *Crap ! Robin Hood airport is closed. You've got a week and a bit to get your shit together otherwise I am blowing the airport sky high !!* ». Cette condamnation a toutefois été annulée en appel (*Chambers v. Director of Public Prosecutions* du 27 juillet 2012, High Court of Justice of England and Wales, EWHC 2157, par. 38).

4 La modération de contenu ne vise pas uniquement la suppression des contenus. D'autres « remèdes » sont également utilisés pour réduire la visibilité ou l'accessibilité du contenu ou le démonétiser (voir E. GOLDMAN, « Content Moderation Remedies », *Michigan Technology Law Review*, 2022, p. 1-59, p. 23 s.).

utilisateurs. Toutefois, les approches des législateurs varient, certains offrant une protection plus large aux plateformes, d'autres, comme le législateur suisse, ne prévoyant pas de régime juridique spécifique en matière de responsabilité indirecte.

Dans un premier temps, nous présenterons le régime de responsabilité indirecte prévu en droit européen et en droit américain. Les législateurs européen et américain ont instauré des régimes d'exemption de responsabilité pour les intermédiaires d'Internet qui stockent et diffusent des contenus publiés par leurs utilisateurs. Ces régimes juridiques préviennent toute responsabilité dans le chef des intermédiaires et leur imposent, à tout le moins indirectement, une obligation de modérer le contenu.

Dans un second temps, nous présenterons les mécanismes de modération mis en place par les réseaux sociaux, ainsi que les voies de recours offertes aux utilisateurs afin de contester les décisions de modération. Notre analyse portera finalement sur l'effectivité de ces mécanismes au regard du droit des utilisateurs à un procès équitable.

II. L'obligation des réseaux sociaux de modérer le contenu publié par leurs utilisateurs

Entre les années 1990 et 2000, les législateurs nationaux (et supranationaux) ont introduit des régimes juridiques de responsabilité des plateformes numériques pour le contenu publié par leurs utilisateurs (responsabilité indirecte)[5]. Les régimes de responsabilité indirecte fixent les conditions auxquelles une plateforme a l'obligation de supprimer ou de rendre inaccessible le contenu publié par ses utilisateurs, autrement dit de modérer ledit contenu. Par conséquent, la question de la responsabilité indirecte est intimement liée à l'obligation des réseaux sociaux de modérer le contenu publié par leurs utilisateurs.

Ne souhaitant pas introduire de réglementation spécifique concernant la responsabilité des plateformes numériques, le législateur suisse estime qu'il est plus judicieux que les tribunaux procèdent à un examen au cas par cas, en

5 J. M. BALKIN, « The Future of Free Expression in a Digital Age », *Pepperdine Law Review*, 2009, p. 427-444, p. 434.

application de la réglementation générale existante[6], notamment les art. 28 CC[7] et 28 CP[8].

Nous limiterons dès lors notre propos aux deux principaux régimes de responsabilité indirecte qui ont été réglementés, à savoir d'une part celui de l'Union européenne (A.) et, d'autre part, celui des États-Unis d'Amérique (B.).

A. Le régime juridique de l'Union européenne

1. Les dérogations en matière de responsabilité

Au sein de l'Union européenne, la responsabilité indirecte des plateformes numériques est régie par la Directive sur le commerce électronique de 2000 (Directive e-commerce)[9]. La Directive e-commerce s'applique aux prestataires de « services de la société de l'information », notion qui englobe des services prestés à distance au moyen d'équipements électroniques de traitement et de stockage des données, à la demande individuelle d'un destinataire de services et, en principe, contre rémunération[10]. Au demeurant, les réseaux sociaux sont rémunérés par le biais du traitement des données personnelles de leurs utilisateurs, notamment à des fins publicitaires[11].

La section 4 du chapitre II de la Directive e-commerce prévoit des dérogations en matière de responsabilité pour les « prestataires intermédiaires »[12]. La notion de prestataires intermédiaires englobe les simples transports au sens de l'art. 12, les fournisseurs de cache au sens de l'art. 13 et les fournisseurs d'hébergement

6 OFCOM, *Rapport sur les intermédiaires et plateformes de communication – Effets sur la communication publique et approches de gouvernance*, 17 novembre 2021, p. 49 (cité : Rapport OFCOM) ; CONSEIL FÉDÉRAL SUISSE, *Rapport sur la responsabilité civile des fournisseurs de services Internet*, 11 décembre 2015, p. 19 et 29 ss ; F. ZELLER, « Art. 28 CP », *in* M. A. NIGGLI/H. WIPRÄCHTIGER (édit.), *Strafrecht I, Basler Kommentar*, vol. II, Bâle 2018, N 103*a*, p. 522.

7 Code civil suisse du 10 décembre 1907 (RS 201).

8 Code pénal suisse du 21 décembre 1937 (RS 311.0).

9 Directive 2000/31/CE du Parlement européen et du Conseil du 8 juin 2000 relative à certains aspects juridiques des services de la société de l'information, et notamment du commerce électronique, dans le marché intérieur (Directive e-commerce ; JO L 178).

10 Art. 1 par. 1 let. *b* de la Directive (UE) 2015/1535 du Parlement européen et du Conseil du 9 septembre 2015 prévoyant une procédure d'information dans le domaine des réglementations techniques et des règles relatives aux services de la société de l'information (JO L 241) ; CJUE C-324/09, *L'Oréal c. eBay* du 12 juillet 2011, point 109.

11 Consid. 18 de la Directive e-commerce.

12 Art. 12 à 15 de la Directive e-commerce. La Directive e-commerce ne régit que les dérogations en matière de responsabilité, en laissant aux États membres le soin de déterminer, dans leur droit interne, à quelles conditions la responsabilité des prestataires intermédiaires peut être engagée.

au sens de l'art. 14, les réseaux sociaux étant qualifiés de fournisseurs d'hébergement au sens de cette dernière disposition[13]. Les fournisseurs d'hébergement délivrent un service consistant à stocker des informations fournies par un destinataire du service (*i.e.* l'utilisateur), à la demande de ce dernier[14].

Au fil de sa jurisprudence, la Cour de justice de l'Union européenne a été amenée à préciser le champ d'application de l'art. 14 de la Directive e-commerce, en appliquant le critère de la *neutralité* du prestataire de services. Pour pouvoir bénéficier des dérogations en matière de responsabilité, un prestataire de services de la société de l'information doit être défini comme « prestataire intermédiaire », en ce sens que son activité est purement technique, automatique et passive[15]. Il convient ainsi d'examiner si le rôle exercé par le prestataire de services est neutre, ce qui implique l'absence de connaissance du contenu qu'il stocke[16]. A contrario, si ledit prestataire joue un rôle actif de nature à lui conférer une connaissance du contenu stocké, il ne peut pas bénéficier des dérogations en matière de responsabilité prévues par la Directive[17].

Selon l'art. 14 par. 1 de la Directive e-commerce, un fournisseur d'hébergement n'est pas responsable des informations stockées à la demande des utilisateurs, à condition que le fournisseur n'ait pas une *connaissance effective* du caractère illicite de l'activité ou de l'information stockée ou, en cas de demande en dommages-intérêts, n'ait pas la connaissance de faits ou de circonstances selon lesquels le caractère illicite de l'activité ou de l'information est *apparent*[18]. En outre, le fournisseur n'est pas responsable lorsque, dès le moment où

13 CJUE C-18/18, *Glawischnig-Piesczek c. Facebook* du 3 octobre 2019, point 22 ; S. STALLA-BOURDILLON, « Des intermédiaires de l'Internet aux plateformes en ligne en passant par les fournisseurs d'hébergement : repenser le paradigme "de la neutralité" à l'aune des droits fondamentaux », *in* J. SÉNÉCHAL/J. ROCHFELD (édit.), *Rôle et responsabilité des opérateurs de plateforme en ligne : approche(s) transversale(s) ou approches sectorielles ? [actes du colloque du 24 novembre 2016]*, Paris 2018, p. 61-86, p. 62 ; P. VAN EECKE, « Online Service Providers and Liability : a Plea for a Balanced Approach », *Common Market Law Review*, 2011, p. 1455-1502, p. 1462 s.

14 Art. 14 par. 1 de la Directive e-commerce.

15 Consid. 42 de la Directive e-commerce ; CJUE C-324/09 *L'Oréal c. eBay* (n. 10), points 112 s.

16 CJUE C-682/18 et C-683/18, *Peterson c. Google* du 22 juin 2021, points 105 s. ; C 324/09 *L'Oréal c. eBay* (n. 10), points 115 s. ; C-236/08 à C-238/08, *Google c. Vuitton* du 23 mars 2010, point 114.

17 CJUE C-682/18 et C-683/18 *Peterson c. Google* (n. 16), point 106 ; C-324/09 *L'Oréal c. eBay* (n. 10), point 113.

18 CJUE C-682/18 et C-683/18 *Peterson c. Google* (n. 16), point 113 ; VAN EECKE (n. 13), p. 1465. En d'autres termes, le caractère illicite de l'activité doit être concrètement établi ou, en cas de demande de dommages-intérêts, aisément indentifiable.

il a connaissance du caractère illicite de l'activité ou de l'information, il agit sans délai (« promptement ») pour supprimer cette dernière[19].

En présence d'une demande de dommages-intérêts, il suffit, pour que le fournisseur d'hébergement soit privé du bénéfice de l'exonération de responsabilité prévue à l'art. 14, que le caractère illicite du contenu soit *apparent*[20]. Sont ainsi visées, par exemple, les situations dans lesquelles l'exploitant d'un réseau social découvre l'existence d'un contenu illicite à la suite d'un examen effectué de sa propre initiative, ainsi que celles dans lesquelles l'existence d'un tel contenu lui est notifiée par un utilisateur[21]. À elle seule, la notification de l'existence d'une information illicite ne saurait toutefois écarter le bénéfice de l'exemption de responsabilité, étant donné que des notifications d'informations prétendument illicites peuvent se révéler erronées ou insuffisamment précises et étayées[22]. En effet, encore faut-il que la notification contienne suffisamment d'éléments pour permettre au prestataire de services de s'assurer, sans examen juridique approfondi, du caractère illicite du contenu[23].

En ce qui concerne les demandes qui n'ont pas pour objet des prétentions en dommages-intérêts, les fournisseurs d'hébergement doivent avoir une connaissance *effective* du caractère illicite pour être privés des dérogations en matière de responsabilité. À cet égard, le simple fait que l'exploitant d'une plateforme de partage de vidéos (en l'occurrence, *YouTube*) mette en œuvre des mesures techniques et automatisées visant à indexer et recommander des vidéos aux utilisateurs, ainsi qu'à détecter le contenu susceptible de porter atteinte aux droits de propriété intellectuelle, n'implique pas que, ce faisant, cet exploitant joue un rôle actif de nature à lui conférer la connaissance concrète du contenu illicite téléversé sur cette plateforme[24].

2. *L'interdiction d'imposer une obligation de surveillance générale*

Dans le cadre de leurs activités, les prestataires de services intermédiaires ne sauraient se voir imposer par les États membres une obligation générale de surveiller les informations qu'ils transmettent ou stockent ou de

19 Le législateur allemand impose aux réseaux sociaux d'agir dans un délai de 24 heures dès le signalement pour supprimer un contenu manifestement illicite (§ 3 ch. 2 du Netzwerkdurchsetzungsgesetz, NetzDG ; BGBl. I, p. 3352). Pour les autres cas de contenus illicites, ce délai est en principe de sept jours (§ 3 ch. 3 NetzDG).

20 CJUE C-324/09, *L'Oréal c. eBay* (n. 10), point 120.

21 CJUE C-682/18 et C-683/18, *Peterson c. Google* (n. 16), point 115.

22 CJUE C-324/09, *L'Oréal c. eBay* (n. 10), point 122.

23 CJUE C-682/18 et C-683/18, *Peterson c. Google* (n. 16), point 116.

24 CJUE C-682/18 et C-683/18, *Peterson c. Google* (n. 16), points 114 à 118.

rechercher activement les faits ou les circonstances relevant d'activités illicites[25]. En effet, une obligation de filtrage systématique des contenus serait disproportionnée et risquerait de pousser les intermédiaires à mettre en place un système de surveillance contraire à la liberté d'expression[26].

L'interdiction pour les États membres de l'Union européenne d'imposer aux prestataires de services une obligation de surveillance ne vaut toutefois que pour les obligations à caractère général[27]. Ainsi, rien n'empêche les autorités étatiques d'imposer aux prestataires des obligations de surveillance applicables à des cas spécifiques ou d'émettre des injonctions permettant de prévenir ou de mettre un terme à des activités illicites[28].

La portée des injonctions émises par les juridictions nationales s'étend aux informations stockées par un fournisseur d'hébergement, dont le contenu est *identique* ou *équivalent* à celui d'une information déclarée au préalable comme étant illicite[29]. En d'autres termes, lorsqu'une juridiction nationale estime, en vertu de son droit interne, qu'une information stockée par un réseau social est illicite, elle peut exiger de celui-ci qu'il supprime non seulement le contenu identique mais aussi le contenu équivalent à l'information déclarée au préalable comme étant illicite[30].

Le critère de l'équivalence vise des informations qui véhiculent un message dont le contenu diverge très peu de celui ayant donné lieu au constat initial d'illicéité[31]. L'injonction d'une juridiction nationale doit ainsi pouvoir s'étendre aux contenus formulés de manière légèrement différente par rapport aux contenus déclarés illicites, en raison notamment des mots employés ou de leur combinaison[32].

25 Art. 15 par. 1 de la Directive e-commerce.

26 CourEDH, *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt c. Hongrie* du 2 février 2016 (requête n° 22947/13), par. 82.

27 Consid. 47 de la Directive e-commerce.

28 Consid. 45 et 47 et art. 14 par. 3 de la Directive e-commerce ; CJUE C-18/18, *Glawischnig-Piesczek* (n. 13), point 25 ; VAN ECKE (n. 13), p. 1464.

29 CJUE C-18/18, *Glawischnig-Piesczek c. Facebook* (n. 13), motifs.

30 CJUE C-18/18, *Glawischnig-Piesczek c. Facebook* (n. 13), motifs. Par ailleurs, il est intéressant de noter que la CJUE accorde aux injonctions nationales une portée extraterritoriale, en ce sens que les juridictions des États membres peuvent obliger un fournisseur d'hébergement à rendre inaccessible des informations illicites au niveau mondial (CJUE C-18/18, *Glawischnig-Piesczek c. Facebook* (n. 13), points 48 à 50).

31 CJUE C-18/18, *Glawischnig-Piesczek c. Facebook* (n. 13), point 39.

32 CJUE C-18/18, *Glawischnig-Piesczek c. Facebook* (n. 13), points 40 s. On pense notamment à des propos diffamatoires publiés sur les réseaux sociaux et qui visent une seule et même personne, dans un contexte bien défini.

3. *Le Règlement sur les services numériques*

En décembre 2020, la Commission européenne a proposé l'adoption d'un Règlement régissant les services numériques (*Digital Services Act*)[33]. Le projet, tel que modifié par le Parlement européen en juillet 2022[34], a été adopté dans sa version finale en octobre 2022[35]. Le Règlement a notamment pour objectif de renforcer les possibilités de recours des utilisateurs de plateformes numériques, en imposant à celles-ci des obligations précises de modération de contenu[36].

À l'instar du Règlement européen sur la protection des données[37], le *Digital Services Act* a un champ d'application extraterritorial, dans la mesure où il s'applique aux services fournis à des citoyens européens, indépendamment du lieu d'établissement des fournisseurs de services[38]. En outre, les services fournis par la plateforme doivent avoir un lien substantiel avec le territoire de l'Union. Un tel lien est réputé exister notamment lorsque le fournisseur de services dirige ses activités vers un ou plusieurs États membres[39].

---

33 COMMISSION EUROPÉENNE, *Proposition de Règlement du Parlement et du Conseil relatif à un marché intérieur des services numériques (Législation sur les services numériques) et modifiant la directive 2000/31/CE*, 15.12.2020 (COM 2020/0361 COD).

34 PARLEMENT EUROPÉEN, *Règlement sur les services numériques – Résolution législative du Parlement européen sur la proposition de Règlement du Parlement Européen et du Conseil relatif à un marché unique des services numériques (Législation sur les services numériques) et modifiant la directive 2000/31/CE*, 05.07.2022 (COM 2020 0825 – C9-0418/2020 – 2020/0361 COD).

35 Règlement (UE) 2022/2065 du Parlement européen et du Conseil du 19 octobre 2022 relatif à un marché unique des services numériques et modifiant la directive 2000/31/CE (DSA ; JO L 277). Dans la suite de cette contribution, les termes de Règlement sur les services numériques et de *Digital Services Act* (DSA) désignent le Règlement dans sa version finale.

36 COMMISSION EUROPÉENNE (n. 33), p. 14.

37 Art. 3 par. 2 du Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (Règlement général sur la protection des données, RGPD ; JO L 119).

38 Art. 2 par. 1 DSA. En ce qui concerne le champ d'application du RGPD, voir L. PAILLER, « L'applicabilité spatiale du Règlement général sur la protection des données (RGPD) – Commentaire de l'article 3 », *Journal du droit international* (Clunet), 2018, p. 823-850, p. 829 et 839 ss.

39 Art. 3 let. *d* et *e* DSA.

En matière de responsabilité indirecte, le Règlement sur les services numériques reprend, dans les grandes lignes, la systématique de la Directive e-commerce[40]. Le Règlement distingue, en tant que fournisseurs de services intermédiaires[41], les simples transports, les fournisseurs de cache et les fournisseurs d'hébergement[42]. Toutefois, un nouveau système dit de « gradation » est introduit par le législateur européen. Ce système prévoit d'imposer des obligations de modération de contenu en fonction du type et de la nature des services proposés par les prestataires intermédiaires. Quatre catégories de fournisseurs de services sont distinguées, de la catégorie la plus large (les prestataires intermédiaires) à la catégorie la plus étroite (les très grandes plateformes en ligne).

Les *prestataires intermédiaires*[43] sont soumis à une obligation de transparence en ce qui concerne, d'une part, les modalités d'application de leurs conditions générales[44] et, d'autre part, leurs activités de modération de contenu[45]. Certaines obligations s'appliquent spécifiquement à la catégorie des *fournisseurs d'hébergement*[46]. Ces derniers devront établir des mécanismes permettant à toute personne de leur notifier, par voie électronique, la présence de contenus considérés comme illicites[47]. Les décisions rendues par les modérateurs devront également être suffisamment motivées, tant factuellement que juridiquement[48].

Le Règlement prévoit, en outre, une série d'obligations spécifiques à charge des *plateformes en ligne*[49] et des *très grandes plateformes en ligne*[50]. Les plateformes en ligne (ou plateformes numériques) sont définies comme des fournisseurs d'hébergement qui, à la demande d'un destinataire du service (*i.e.* l'utilisateur), stockent des informations et les diffusent au public[51]. Les très grandes plateformes en ligne sont des plateformes en ligne qui ont une incidence sociétale et économique particulière. Au sens du Règlement, une plateforme en ligne

40 Chapitre II, DSA. Les art. 12 à 15 de la Directive e-commerce seront supprimés et remplacés par les art. 4 à 6 et 8 du Règlement sur les services numériques (art. 89 DSA).

41 La Directive e-commerce utilise la notion de « prestataires intermédiaires » (voir *supra*, A.1.).

42 Art. 4 à 6 DSA.

43 Section 1, Chapitre III, DSA.

44 Art. 14 par. 1 DSA.

45 Selon l'art. 15 par. 1 DSA, les fournisseurs d'hébergement doivent établir, au moins une fois par an, un rapport clair et facilement compréhensible sur leurs activités de modération de contenu.

46 Section 2, Chapitre III, DSA.

47 Art. 16 par. 1 et 2 DSA.

48 Art. 17 et 20 par. 5 DSA

49 Sections 2 et 3, Chapitre III, DSA.

50 Section 5, Chapitre III, DSA.

51 Art. 3 let. *i* DSA et consid. 13 DSA. Par « diffusion au public », il faut entendre le fait de mettre des informations à la disposition d'un nombre potentiellement illimité de tiers, à la demande du destinataire du service et sans l'intervention de ce dernier (art. 3 let. *k* DSA ; consid. 14 DSA).

est « très grande » lorsqu'elle fournit mensuellement ses services à au moins 45 millions d'utilisateurs actifs au sein de l'Union[52]. Les très grandes plateformes en ligne auront l'obligation de lutter contre les risques systémiques, tels que la diffusion de contenus illicites, ainsi que tout effet négatif sur le discours civique ou le processus électoral[53].

B. Le régime juridique des États-Unis d'Amérique

1. L'immunité des plateformes : la section 230 du Communications Decency Act

Afin d'engager la responsabilité des intermédiaires d'Internet, les tribunaux américains ont longtemps appliqué le critère du contrôle éditorial[54]. En effet, l'intermédiaire qui exerçait un contrôle éditorial sur le contenu publié par autrui était considéré comme l'éditeur (ou le *publisher*) d'une publication[55]. En d'autres termes, les intermédiaires qui mettaient en place des mécanismes de modération de contenu étaient considérés comme responsables du contenu illicite publié par autrui sur leur site[56].

Le critère du contrôle éditorial s'est rapidement révélé être problématique, dans la mesure où il assimile les intermédiaires en ligne à des médias traditionnels (journalisme, radio et télévision). Ces derniers exercent effectivement un contrôle éditorial de l'information publiée, qui se doit d'être qualitative, objective et véridique, dans un but d'utilité sociale[57]. Les intermédiaires d'Internet

52 Art. 33 par. 1 DSA.

53 Art. 34 par. 1 DSA.

54 R. D. BASTIAN, « Content Moderation Issues Online : Section 230 Is Not to Blame », *Texas A&M Journal of Property Law*, 2022, p. 43-72, p. 49 s. ; K. KLONICK, « The New Governors : The People, Rules, and Processes Governing Online Speech », *Harvard Law Review*, p. 1598-1670, p. 1604 ; T. ROMANOFF, *The Future of Intermediary Liability and Content Moderation, Bipartisan Policy Center*, 16.03.2022, https://bipartisanpolicy.org/blog/the-future-of-intermediary-liability-and-content-moderation/, consulté le 31.08.2022.

55 *Stratton-Oakmont v. Prodigy Servs. Co.* du 26 mai 1995, US Supreme Court of New York (1995 N.Y. Misc. LEXIS 229), par. 10 s.

56 *Stratton-Oakmont v. Prodigy Servs.* (n. 55), par. 10 s. ; *Cubby, Inc. v. Compuserve, Inc.* du 29 octobre 1991, US District Court for the Southern District of New York (776 F. Supp. 135), par. 140, dans lequel le tribunal n'a pas considéré comme responsable du contenu illicite publié par des tiers le simple « distributeur » d'une information ; KLONICK (n. 54), p. 1605.

57 En Suisse, l'activité journalistique est régie par un Code de déontologie, qui impose notamment une obligation de véracité de l'information (CONSEIL SUISSE DE LA PRESSE, *Droits et devoirs du/de la journaliste*, https://presserat.ch/fr/code-de-deontologie-des-journalistes/erklaerungen/, consulté le 31.08.2022). En matière de radio et télévision, l'art. 4 al. 1 et 2 de la loi fédérale sur la radio et la télévision (LRTV ; RS 784.40)

publient, quant à eux, des informations générées par leurs utilisateurs[58]. Par conséquent, conditionner la responsabilité des intermédiaires d'Internet au critère du contrôle éditorial risque soit de les dissuader de modérer le contenu publié par leurs utilisateurs, soit d'engendrer une censure caractérisée par une suppression excessive dudit contenu, ceci dans le but d'échapper à toute forme de responsabilité[59].

Afin de remédier à l'insécurité juridique générée par le critère du contrôle éditorial, le Congrès américain a adopté la section 230 du *Communications Decency Act* (CDA) en 1996[60]. La section 230 CDA et son interprétation par les tribunaux américains ont largement façonné la liberté d'expression sur Internet[61]. En effet, cette loi représente le « socle » sur la base duquel les géants du numérique ont construit et développé leur politique et leur système de modération de contenu[62].

Le droit américain garantit aux plateformes numériques une large immunité pour le contenu publié par leurs utilisateurs[63]. Souvent qualifiée de « loi fondamentale pour la parole sur le Web »[64], la section 230 CDA traduit la conception libérale des États-Unis en matière de liberté d'expression, protégée par le Premier Amendement de la Constitution[65]. L'importance de la liberté d'expression

souligne que les émissions doivent respecter les droits fondamentaux et représenter les événements de manière fidèle.

58 E. DOUEK, *Content Moderation as System Thinking*, 2022, p. 5, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4005326, consulté le 31.08.2022 (à paraître dans la *Harvard Law Review*) ; J. C. YORK/E. ZUCKERMAN, « Moderating the Public Sphere », *in* R. F. JØRGENSEN, *Human Rights in the Age of Platforms*, Cambrige/Londres 2019, p. 137-161, p. 137.

59 BALKIN (n. 5), p. 436 ; S. M. CORDERO, « Damnum Absque Injuria : Zeran v. AOL and Cyberspace Defamation Law », *Fordham Intellectual Property, Media and Entertainment Law Journal*, 1999, p. 775-882, p. 792 ; KLONICK (n. 54), p. 1605 et 1608.

60 US Code, Titre 47, § 230 du 8 février 1996 (Pub. L. 104-104 ; cité : 47 USC § 230 CDA) ; CORDERO (n. 59), p. 795, citant le Congrès américain : « [O]ne of the specific purposes of [section 230] is to overrule [past decisions] which have treated such providers and users as [p]ublishers or speakers of content that is not their own because they have restricted access to objectionable material ».

61 BASTIAN (n. 54), p. 49 ; KLONICK (n. 54), p. 1604.

62 ROMANOFF (n. 54).

63 *Doe v. Backpage.com, LLC* du 14 mars 2016, US Court of Appeals for the First Circuit (817 F.3d 12), par. 18 ; *Universal Commun. Sys. v. Lycos, Inc.* du 23 février 2007, US Court of Appeals for the First Circuit (478 F.3d 413), par. 419 ; E. B. LAIDLAW, « What Is a Joke ? – Mapping the Path of a Speech Complaint on Social Networks », *in* D. MANGAN/L. E. GILLIES (édit.), *The Legal Challenges of Social Media*, Cheltenham/Northampton 2017, p. 127-154, p. 139 s.

64 KLONICK (n. 54), p. 1604.

65 Sur la conception libérale de la liberté d'expression aux États-Unis, voir A. KOLTAY, *New media and freedom of expression : rethinking the constitutional foundations of the public sphere*, Oxford/Chicago 2019, p. 24 s.

a pour corollaire une certaine hostilité à l'égard de toute forme de régulation du contenu diffusé sur Internet[66].

Selon la section 230 CDA, « *no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider* »[67]. Dès lors que l'activité de l'exploitant implique des fonctions éditoriales dites « traditionnelles », celui-ci est immunisé pour le contenu publié par ses utilisateurs[68]. En revanche, lorsque l'opérateur génère, contrôle ou modifie activement le contenu publié sur son site, l'immunité cesse de trouver application[69].

Les tribunaux américains ont interprété la section 230 CDA de façon large et ambiguë[70], accordant une immunité aux intermédiaires même lorsque ceux-ci facilitent la diffusion de contenu à caractère pédopornographique[71] ou le trafic sexuel d'êtres humains[72]. Dans l'affaire *MySpace*, l'exploitant d'un réseau social a été considéré comme n'étant pas responsable du fait que, sur son site, des prédateurs sexuels puissent entrer en contact avec des personnes mineures et commettre des infractions sexuelles[73]. En revanche, a été considéré comme responsable de discriminations, l'exploitant d'un site de recherche de colocations qui permet aux utilisateurs de rechercher des colocataires en fonction de leurs caractéristiques personnelles (couleur de peau, sexe, orientation sexuelle, etc.)[74]. Selon la Cour d'appel en charge de l'affaire, l'opérateur en question était suffisamment impliqué dans la conception et les opérations du moteur de recherche et du système d'email de la plateforme – conçue pour limiter l'accès

66 M. MACCARTHY, « A Consumer Protection Approach to Platform Content Moderation in the United States », *in* B. PETKOVA/T. OJANEN (édit.), *Fundamental Rights Protection Online*, Cheltenham/Northampton 2020, p. 115-139, p. 116.

67 47 USC § 230 (c) (1).

68 *Doe v. Backpage.com, LLC* (n. 63), par. 21 ; *Universal Commun. Sys. v. Lycos, Inc.* (n. 63), par. 422 ; *Zeran v. America Online, Inc.* du 12 novembre 1997, US Court of Appeals for the Fourth Circuit (129 F.3d 327), par. 330 ; KOLTAY (n. 65), p. 99.

69 KOLTAY (n. 65), p. 164 ; LAIDLAW (n. 63), p. 140.

70 *Malvarebytes, Inc. v. Enigma Software Grp. USA, LLC* du 13 octobre 2020, US Supreme Court (141 S. Ct. 13), par. 15 ; KOLTAY (n. 65), p. 99 ; LAIDLAW (n. 63), p. 140, n. 66, qui souligne que « [t]he line between [traditional] acceptable editing content and editing content that loses the immunity is not clear, although the Electronic Frontier Foundation (EFF) advises the focus is likely on whether the edit changes the meaning of the information [...] ».

71 *Doe v. America Online* du 8 mars 2001, US Supreme Court of Florida (783 So. 2d 1010), par. 1018.

72 *Doe v. Backpage.com, LLC* (n. 63). En 2017, le législateur a toutefois modifié la section 230 CDA en introduisant une exemption d'immunité pour les cas de trafic sexuel d'êtres humains (47 USC § 230 (e) (5)).

73 *Doe v. MySpace Inc.* du 16 mai 2008, US Court of Appeals for the Fifth Circuit (528 F.3d 413), par. 420 à 422.

74 *Fair Housing Council of San Fernando Valley v. Roommates.com LLC* du 3 avril 2008, US Court of Appeals for the Ninth Circuit (521 F.3d 1157).

au logement sur la base de caractéristiques protégées – pour ne pas être au bénéfice de l'immunité garantie par la section 230 CDA[75].

Contrairement au droit européen[76], le fait que l'intermédiaire possède une connaissance effective ou apparente du contenu illicite publié sur sa plaforme n'exclut pas l'application de l'immunité prévue par le *Communications Decency Act*[77]. En effet, le fait d'imposer aux plateformes numériques une responsabilité dépendant de la connaissance du caractère illicite du contenu entraînerait une censure sur Internet[78]. Les réseaux sociaux seraient potentiellement responsables à chaque fois qu'ils recevraient un signalement de la part des utilisateurs. Chaque signalement demanderait un examen minutieux et rapide des circonstances qui entourent la publication, un examen juridique du caractère illicite de cette dernière et une décision éditoriale qui risquerait d'entraîner la responsabilité de la plateforme, dans le cas où le contenu se révèlerait être effectivement illicite[79].

La portée de l'immunité garantie par la section 230 CDA n'est cependant pas illimitée[80]. En effet, l'immunité ne s'applique pas à des contenus qui violent le droit pénal fédéral, le droit relatif à la propriété intellectuelle et la législation sectorielle des différents États américains[81].

2. *La clause du « bon samaritain »*

La section 230 CDA prévoit une protection contre toute forme de responsabilité pour les « bons samaritains », c'est-à-dire les intermédiaires qui recherchent activement et suppriment le contenu illicite et indésirable publié par leurs utilisateurs[82]. En effet, les fournisseurs de services en ligne (*interactive computer service*) ne doivent pas être tenus pour responsables des actions de

75 *Id.*, par. 1170.
76 Art. 14 par. 1 de la Directive e-commerce (voir *supra*, II.A.1.).
77 KOLTAY (n. 65), p. 99 ; LAIDLAW (n. 63), p. 140.
78 *Zeran v. America Online, Inc* (n. 68), par. 333 : « Because service providers would be subject to liability only for the publication of information, and not for its removal, they would have a natural incentive simply to remove messages upon notification, whether the contents were defamatory or not. Thus, like strict liability, liability upon notice has a chilling effect on the freedom of Internet speech ».
79 *Zeran v. America Online, Inc.* (n. 68), par. 333.
80 LAIDLAW (n. 63), p. 140.
81 47 USC § 230 (e) CDA ; LAIDLAW (n. 63), p. 140. En matière de propriété intellectuelle, l'exception d'immunité vaut autant pour le droit fédéral que pour le droit des États américains (47 USC § 230 (e) (2) CDA ; *Hepp v. Facebook* du 23 septembre 2021, US Court of Appeals for the Third Circuit, 14 F.4th 204).
82 KLONICK (n. 54), p. 1605.

modération entreprises pour restreindre l'accès à des contenus obscènes, violents, haineux ou autrement répréhensibles[83].

Le Congrès américain a adopté la clause du bon samaritain dans le but d'encourager l'autorégulation de la parole en ligne par les plateformes numériques[84]. Ces dernières sont incitées par le législateur à utiliser des mécanismes de modération, sans craindre d'engager leur responsabilité[85]. L'objectif est d'inciter la suppression des contenus illicites, tout en protégeant la liberté d'expression des utilisateurs sur Internet[86].

Le droit de modérer le contenu généré par ses utilisateurs constitue un droit fondamental de la plateforme, protégé par le Premier Amendement de la Constitution et la section 230 CDA[87]. Les plateformes numériques restent toutefois libres de mettre en place des mécanismes de modération de contenu et de supprimer les contenus illicites[88].

En pratique, la majorité des réseaux sociaux ont mis en place des mécanismes de modération de contenu afin d'éviter que leurs utilisateurs ne soient confrontés à des contenus illicites ou choquants. Ces mécanismes combinent les outils technologiques et l'expertise humaine afin de détecter certains contenus et d'apprécier leur caractère illicite ou indésirable.

III. Les mécanismes de modération de contenu

Au niveau temporel, on peut différencier la modération *ex ante* ou algorithmique (A.) de la modération *ex post* ou humaine (B.). La modération *ex ante* intervient avant que le contenu ne soit publié sur la plateforme, tandis que la modération *ex post* intervient après la publication[89]. Tant la modération

83 47 USC § 230 (c) (2) (A).

84 *Batzel v. Smith* du 24 juin 2003, United States Court of Appeals for the Ninth Circuit (333 F. 3d 1018), par. 1028 ; KLONICK (n. 54), p. 1607.

85 *Doe v. Backpage.com LLC* (n. 63), par. 19 ; BASTIAN (n. 54), p. 50.

86 47 USC § 230 (b) (4) : « It is the policy of the United States [...] to remove disincentives for the development and utilization of blocking and filtering technologies that empower parents to restrict their children's access to objectionable or inappropriate online material » ; *Zeran v. America Online, Inc.* (n. 68), par. 331.

87 *NetChoice, LLC v. Paxton* du 31 mai 2022, US Supreme Court (142 S. Ct. 1715) ; *Manhattan Cmty. Access Corp v. Halleck* du 17 juin 2019, US Supreme Court (139 S. Ct. 1921), par. 1932 (reconnaissant le droit de certaines entités privées d'exercer un contrôle éditorial sur le contenu publié par autrui sur leur plateforme) ; A. CALLAMARD, « The Human Rights Obligations of Non-State Actors », *in* R. F. JØRGENSEN, *Human Rights in the Age of Platforms*, Cambrige/Londres 2019, p. 191-225, p. 208.

88 BASTIAN (n. 54), p. 53.

89 GOLDMAN (n. 4), p. 51 ; KLONICK (n. 54), p. 1635.

ex ante que la modération *ex post* peuvent être proactives ou réactives[90]. Dans le premier cas, des algorithmes ou des équipes spécialisées de modérateurs recherchent activement les contenus illicites (par exemple, en matière de terrorisme)[91] ou contraires aux conditions générales de la plateforme (par exemple, la nudité)[92]. Dans le second cas, un algorithme ou un utilisateur signale le contenu aux modérateurs humains, à charge pour ces derniers d'en déterminer le caractère illicite ou contraire aux conditions générales[93].

Les mécanismes de modération de contenu doivent respecter certains standards procéduraux afin d'être effectifs (C.). Les réseaux sociaux garantissent notamment à leurs utilisateurs la possibilité de contester les décisions de modération de contenu. Le futur Règlement sur les services numériques va encore plus loin, en imposant aux plateformes numériques des obligations précises en matière de modération.

A. La modération de contenu automatisée

1. Le rôle des algorithmes

Les réseaux sociaux ont recours à des moyens automatisés afin de repérer, traquer et supprimer les contenus illicites[94]. À vrai dire, la majorité du contenu illicite et contraire aux conditions générales est repérée par des moyens automatisés, sans qu'un modérateur humain n'en prenne connaissance[95]. Selon le groupe *Meta*, la technologie permettrait de détecter, dans plus de 90 % des cas, le caractère illicite d'un contenu[96].

90 KLONICK (n. 54), p. 1635.

91 D. MADHOK, *How Social Media Is Dealing with the Taliban Takeover*, CNN, www.cnn.com/2021/08/17/tech/facebook-twitter-taliban-hnk-intl/index.html, consulté le 31.08.2022 ; KLONICK (n. 54), p. 1638.

92 META, *Standards de la communauté Facebook – Nudité et activités sexuelles chez les adultes*, https://transparency.fb.com/fr-fr/policies/community-standards/adult-nudity-sexual-activity/, consulté le 31.08.2022.

93 KLONICK (n. 54), p. 1638.

94 H. BLOCH-WEHBA, « Automation in Moderation », *Cornell International Law Journal*, 2020, p. 41-96, p. 41 ; K. SWISHER, « Zuckerberg : The Recode Interview », *Vox*, 18.07.2018, www.vox.com/2018/7/18/17575156/mark-zuckerberg-interview-facebook-recode-kara-swisher, consulté le 31.08.2022.

95 KLONICK (n. 54), p. 1636.

96 META, *Comment la technologie détecte-t-elle les infractions ?*, https://transparency.fb.com/fr-fr/enforcement/detecting-violations/technology-detects-violations/, consulté le 31.08.2022.

Les systèmes automatisés sont justifiés par la quantité et la variété des contenus diffusés sur les réseaux sociaux, ainsi que par la rapidité à laquelle l'information se propage sur Internet[97]. Afin de modérer l'océan de contenu publié sur les réseaux sociaux, les exploitants des plateformes ont adopté une approche industrielle de la modération de contenu, qui crée une sorte « d'usine des décisions »[98]. En effet, l'utilisation de méthodes d'intelligence artificielle (IA) permet de détecter et bloquer un contenu illicite avant même que celui-ci ne soit publié sur la plateforme[99]. La modération automatisée a ainsi lieu entre le téléchargement du contenu et sa publication (modération de contenu *ex ante*)[100]. De plus, l'IA identifie les contenus les plus choquants, comme les traitements dégradants et inhumains, les contenus à caractère pédopornographique, les meurtres, les suicides et les viols. Ainsi, les modérateurs humains sont moins confrontés à ce type de contenus et peuvent se concentrer sur l'analyse du contenu moins dommageable pour leur état de santé[101].

Les réseaux sociaux utilisent des techniques issues des statistiques et de l'informatique pour analyser le caractère illicite des informations et, cas échéant, supprimer ces dernières[102]. Dans le cadre de cette contribution, nous nous limiterons à analyser deux méthodes de modération automatisée, la correspondance (*matching*) et le *machine learning*.

2. *La correspondance*

La correspondance (*matching*) consiste à comparer l'empreinte numérique de deux documents[103]. Cette méthode implique un « hachage » de la publication analysée, à savoir un processus cryptographique qui permet, par le

97 N. APPELMAN *et al.*, *Access to Digital Justice : In Search of an Effective Remedy for Removing Unlawful Online Content*, 2021, p. 9, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3961390, consulté le 31.08.2022 (à paraître dans : X. KRAMER *et al.* (édit.), *Frontiers in Civil Justice*, Cheltenham/Northampton 2022) ; T. GILLESPIE, « Content Moderation, AI, and the Question of Scale », *Big Data & Society*, 2020, p. 1.

98 R. CAPLAN, « Content or Context Moderation : Artisanal, Community-Reliant, and Industrial Approaches », *Data & Society*, 2020, p. 24 ; DOUEK (n. 56), p. 11.

99 GILLESPIE (n. 97), p. 1.

100 KLONICK (n. 54), p. 1636.

101 GILLESPIE (n. 97), p. 4.

102 R. GORWA/R. BINNS/C. KATZENBACH, « Algorithmic Content Moderation : Technical and Political Challenges in the Automation of Platform Governance », *Big Data & Society* 2020, p. 3.

103 GILLESPIE (n. 97), p. 4, qui décrit cette méthode comme la comparaison d'un nouveau contenu avec des exemples de contenus illicites conservés dans une liste noire.

biais d'une fonction de *hash*, d'attribuer à la publication une empreinte numérique (un *hash*)[104]. Une fonction de *hash* est une fonction mathématique qui permet de représenter une donnée – un *tweet*, une image, un document – en une valeur cryptographique unique et propre à cette donnée[105].

De nombreux réseaux sociaux utilisent des méthodes de *matching*, notamment pour lutter contre le contenu à caractère pornographique ou le terrorisme[106]. Le programme *PhotoDNA*, développé par *Microsoft*, permet de comparer les valeurs de *hash* d'images et de vidéos publiées et téléchargées par des utilisateurs avec une base de données qui contient des valeurs de *hash* d'images et de vidéos à caractère pédopornographique[107]. Si le *hash* du contenu en question correspond aux *hash* de la base de données, il s'agit d'un contenu à caractère pédopornographique[108].

La plateforme *YouTube* a développé *Content ID*, un système automatisé d'identification des contenus protégés par le droit d'auteur[109]. Grâce à ce système, le titulaire d'un droit d'auteur détient une empreinte numérique du contenu protégé par son droit. Lorsque le système identifie une correspondance entre l'empreinte numérique d'une vidéo litigieuse et l'empreinte détenue par le titulaire du droit d'auteur, ce dernier peut choisir de bloquer la vidéo, de la monétiser ou de suivre ses statistiques de visionnage[110].

La correspondance est une méthode efficace pour détecter les contenus répétitifs et identiques. Toutefois, en raison de leur sensibilité aux altérations, les fonctions de *hash* basées sur la cryptographie sont peu utiles en matière de modération de contenu[111]. En effet, une manipulation minime du contenu analysé (par exemple, dans la couleur des pixels d'une image) peut modifier le *hash*

104 N. DUARTE/E. LLANSO/A. LOUP, *Mixed Messages ? The Limits of Automated Social Media Content Analysis*, 2017, p. 9 ; GORWA/BINNS/KATZENBACH (n. 102), p. 4 ; YORK/ZUCKERMAN (n. 58), p. 150.

105 DUARTE/LLANSO/LOUP (n. 102), p. 9 ; GORWA/BINNS/KATZENBACH (n. 102), p. 4.

106 BLOCH-WEHBA (n. 94), p. 57 s. ; GILLESPIE (n. 97), p. 3 ; H. WESTERMANN, *An Interactive Guide to the Blockchain, Part II – Hash Functions*, *Laboratoire de cyberjustice*, www.cyberjustice.ca/2020/05/26/an-interactive-guide-to-the-blockchain-part-ii-hash-functions/, consulté le 31.08.2022.

107 BLOCH-WEHBA (n. 94), p. 58 ; MICROSOFT, *PhotoDNA*, www.microsoft.com/en-us/photodna, consulté le 31.08.2022.

108 BLOCH-WEHBA (n. 94), p. 58.

109 YOUTUBE, *Fonctionnement de Content ID*, https://support.google.com/youtube/answer/2797370?hl=fr, consulté le 31.08.2022 ; voir ég. BLOCH-WEHBA (n. 94), p. 64.

110 YOUTUBE (n. 107). La (dé)monétisation vise à partager, entre le titulaire d'un droit d'auteur et l'auteur d'une publication illicite, les gains économiques liés au visionnage de celle-ci.

111 GORWA/BINNS/KATZENBACH (n. 102), p. 4.

dudit contenu et ainsi altérer la correspondance entre celui-ci et le *hash* du contenu illicite, identifié au préalable[112]. Pour cette raison, des méthodes non cryptographiques sont utilisées et visent à calculer les « similarités » entre deux informations, plutôt que de véritables correspondances[113].

3. *Le machine learning*

Alors que la correspondance implique la collecte et la conservation, dans une base de données, des valeurs de *hash* correspondant à des exemples individuels de contenus illicites, la classification implique de généraliser les caractéristiques d'un set de données, au sein duquel des contenus futurs peuvent être classés automatiquement, par « prédiction »[114].

La classification implique des méthodes de *machine learning*. Par « apprentissage » (*learning*), on entend un traitement inductif qui utilise des approches statistiques pour construire, à partir d'un set de données, des modèles qui peuvent réaliser des « prédictions » avec de nouvelles données d'entrée[115]. Dans l'une de ses contributions, ÉRIC TALLEY décrit, en six étapes, une méthode de classification dans le cadre de l'analyse de documents[116]. L'objectif principal de cette méthode est l'extraction d'entrées (*inputs*) parmi un set de données (des documents), dans l'objectif d'arriver à un certain résultat (*output*)[117]. Ces données sont ensuite « purifiées »[118], puis analysées par le biais de méthodes statistiques et de calculs de probabilités[119].

Le *machine learning* utilisé par les réseaux sociaux pour modérer le contenu est, en principe, « supervisé ». On parle d'apprentissage supervisé, dans le sens où un set de données préexistant est labélisé dans différentes catégories (par exemple, discours haineux ou discours licite) afin d'entraîner l'algorithme à

112 GORWA/BINNS/KATZENBACH (n. 102), p. 4.
113 *Ibid.*
114 GORWA/BINNS/KATZENBACH (n. 102), p. 5.
115 GORWA/BINNS/KATZENBACH (n. 102), p. 5 ; Y. MENECEUR/C. BARBARO, « Artificial Intelligence and the Judicial Memory : The Great Misunderstanding », *AI and Ethics*, 2022, p. 260-275. p. 270.
116 E. L. TALLEY, « Is the Future of Law a Driverless Car ? : Assessing How the Data-Analytics Revolution Will Transform Legal Practice », *Journal of Institutional and Theoretical Economics*, 2018, p. 183-205, p. 189 ss.
117 TALLEY (n. 116), p. 189 s.
118 TALLEY (n. 116), p. 190. La ponctuation (virgules, points d'exclamation, etc.) est supprimée. Les termes qui se ressemblent sont réduits en un terme commun (par exemple, les termes *walking*, *walked*, *walks* et *walkable* sont réduits au terme *walk*).
119 TALLEY (n. 116), p. 190 ss.

« apprendre » quelles caractéristiques amènent à classifier une information dans l'une ou l'autre des catégories[120].

Selon le type de contenu en cause et la précision de la prédiction opérée par la technologie, l'algorithme n'a qu'une fonction de filtre, en ce sens que la décision de modération revient *in fine* aux modérateurs humains[121]. En effet, dans la plupart des cas, l'algorithme n'est qu'un soutien à la prise de décision, c'est-à-dire un moyen de détection du contenu illicite[122]. Dans un futur proche, il n'est toutefois pas exclu que la technologie puisse rendre des décisions de modération sans l'intervention d'un être humain[123].

B. Les modérateurs humains

1. Les signalements des utilisateurs

Les utilisateurs d'un réseau social sont de véritables « filtres » de l'information. Bien que jugés par certains comme inefficaces[124], les signalements (*flagging*) jouent un rôle fondamental dans l'activité de modération de contenu[125]. Un signalement est un mécanisme qui permet à un utilisateur d'exprimer son avis au sujet d'un contenu qu'il estime être illicite ou contraire aux conditions générales[126].

120 GORWA/BINNS/KATZENBACH (n. 102), p. 5.

121 YORK/ZUCKERMAN (n. 58), p. 157.

122 K. KLONICK, « The Facebook Oversight Board : Creating an Independent Institution to Adjudicate Online Free Expression », *The Yale Law Journal*, 2021, p. 2418-2499, p. 2431 ; GILLESPIE (n. 97), p. 4 ; YORK/ZUCKERMAN (n. 58), p. 150.

123 META, *Comment les solutions de contrôle fonctionnent-elles ?*, https://transparency.fb.com/fr-fr/enforcement/detecting-violations/how-enforcement-technology-works/, consulté le 31.08.2022 ; plus nuancé : A. HUQ, « A Right to a Human Decision », *Virginia Law Review*, 2020, p. 611-688.

124 COMMISSION EUROPÉENNE (n. 33), p. 10, qui relève qu'une grande partie des utilisateurs qui notifient des contenus illégaux aux fournisseurs de services numériques sont insatisfaits de la réponse apportée et de l'inefficacité des mécanismes de notification ; HATEAID, *Transparenzberichte : Wie Social-Media-Plattformen unsere Rechte umgehen*, 2022, https://hateaid.org/transparenzberichte-social-media-plattformen, consulté le 31.08.2022. Entre 8 et 15 % du contenu signalé en Allemagne est effectivement supprimé par les plateformes de réseaux sociaux.

125 KLONICK (n. 122), p. 2432.

126 KLONICK (n. 54), p. 1638 ; A. KUCZERAWY, « From "Notice and Takedown" to "Notice and Stay Down" : Risks and Safeguards for Freedom of Expression », *in* G. FROSIO (édit.), *Oxford Handbook of Online Intermediary Liability*, Oxford 2020, p. 524-543, p. 528.

Des millions de contenus sont signalés chaque jour sur *Facebook*[127]. En principe, le contenu signalé par les utilisateurs est ensuite stocké sur un serveur, dans l'attente d'être analysé par des modérateurs humains[128].

L'adoption d'un système de signalement permet à la plateforme d'externaliser ses activités, ce qui induit une véritable dilution du travail de détection des activités illicites[129]. En effet, les signalements sont efficaces pour modérer un grand nombre de publications et ainsi réduire la charge de travail de la plateforme[130]. De plus, les réseaux sociaux coopèrent avec des « signaleurs de confiance » (*trusted flaggers*), à savoir des utilisateurs privés, des autorités publiques ou des ONG qui signalent des contenus illicites ou indésirables aux plateformes[131]. Par exemple, en Suisse, certains réseaux sociaux accordent le statut de *trusted flaggers* à Office fédéral de la police (fedpol)[132].

Par ailleurs, certains régimes juridiques sous-tendent la responsabilité indirecte des plateformes numériques à la connaissance du contenu illicite, en introduisant des systèmes de « *notice and takedown* » (voir *supra*, II.)[133]. En droit européen, les signalements des utilisateurs sont présumés donner lieu à la connaissance du caractère illicite d'un contenu, lorsque ce caractère est apparent[134]. En d'autres termes, lorsque le signalement permet aux modérateurs de constater, sans examen juridique approfondi, le caractère illicite du contenu signalé, l'exploitant du réseau social est présumé avoir la connaissance du contenu illicite et doit supprimer ce dernier, ceci afin de pouvoir bénéficier des exemptions de responsabilité prévues par la Directive e-commerce[135]. À défaut, le réseau social pourra être tenu pour responsable du contenu illicite publié par l'utilisa-

127 FACEBOOK, *Signaler un abus*, www.facebook.com/help/1380418588640631, consulté le 31.08.2022 ; KLONICK (n. 122), p. 2432.

128 KLONICK (n. 122), p. 2433 ; KLONICK (n. 54), p. 1639 ; YORK/ZUCKERMAN (n. 58), p. 149, qui évoquent un travail de l'ombre du modérateur.

129 En juin 2022, presque trois milliards d'utilisateurs étaient actifs sur *Facebook* mensuellement, ce qui représente une force de signalement considérable (voir META, *Meta Reports Second Quarter 2022 Results*, juillet 2022, https://investor.fb.com/investor-news/press-release-details/2022/Meta-Reports-Second-Quarter-2022-Results/default.aspx, consulté le 31.08.2022.

130 KLONICK (n. 54), p. 1638.

131 Consid. 61 et art. 22 DSA.

132 Rapport de l'OFCOM (n. 6), p. 55.

133 KUCZERAWY (n. 126), p. 526.

134 Art. 16 par. 3 DSA, qui codifie la jurisprudence de la CJUE C-682/18 et C-683/18, *Peterson c. Google* (n. 16).

135 Art. 14 par. 1 de la Directive e-commerce et art. 6 DSA.

teur, en application du droit interne des États membres de l'Union européenne[136]. En matière de droit de propriété intellectuelle, le législateur américain a mis en place un régime similaire[137]. Ce système permet aux titulaires de droit de propriété intellectuelle de signaler à l'exploitant du réseau social toute utilisation indue de leur droit. Si l'exploitant supprime rapidement le contenu signalé, il est exempté de toute responsabilité[138].

Les mécanismes permettant aux utilisateurs de signaler un contenu devraient être faciles d'accès et d'utilisation. Ils devraient notamment permettre aux utilisateurs de notifier aisément au réseau social les éléments d'information spécifiques permettant à ce dernier de déterminer le caractère illicite du contenu en question[139].

2. *Les équipes chargées de modérer le contenu*

Les réseaux sociaux externalisent leurs activités de modération de contenu dans le monde entier, dans ce qui ressemble à de véritables *call centers*[140]. *Meta* emploie par exemple plus de 15 000 modérateurs[141]. Les modérateurs sont de véritables « fourmis ouvrières » de la régulation de la parole en ligne, travaillant dans l'ombre et dans des conditions extrêmement difficiles[142].

L'activité des modérateurs consiste à interpréter et appliquer les conditions générales du réseau social relatives à la modération de contenu, auxquelles les utilisateurs sont liés contractuellement[143].

---

136 Au sujet du régime de responsabilité indirecte prévu par la Directive e-commerce et le Règlement sur les services numériques, voir *supra*, II.A.

137 Section 512 du *Digital Millennium Copyright Act* (DMCA) du 28 octobre 1998 (H.R. 2281, 112 Stat. 2860 ; cité : § 512 DMCA) ; LAIDLAW (n. 63), p. 140.

138 § 512 (c) (1) (C) DMCA : « A service provider shall not be liable for infringement of copyright […], if the service provider upon notification of claimed infringement […], responds expeditiously to remove, or disable access to, the material that is claimed to be infringing or to be the subject of infringing activity ».

139 Consid. 50 et art. 16 DSA.

140 KLONICK (n. 54), p. 1640.

141 META, *Détecter les infractions*, https://transparency.fb.com/fr-fr/enforcement/detecting-violations/, consulté le 31.08.2022.

142 En effet, les modérateurs sont confrontés à des contenus choquants, qui peuvent engendrer un stress psychologique important (voir S. ROBERTS, *Behind the Screen : Content Moderation in the Shadows of Social Media*, New York 2019 ; RTSINFO, *Les « nettoyeurs » de Facebook, des travailleurs sous pression*, 2020, www.rts.ch/info/sciences-tech/11070120-les-nettoyeurs-de-facebook-des-travailleurs-sous-pression.html, consulté le 31.08.2022).

143 META, *Standards de la communauté Facebook*, https://transparency.fb.com/fr-fr/policies/community-standards, consulté le 31.08.2022 ; KOLTAY (n. 65), p. 183.

La modération de contenu est un exercice complexe[144]. En effet, il est souvent difficile de déterminer le caractère illicite ou la contrariété aux conditions générales de certains contenus, tels que les discours violents ou haineux et la désinformation[145]. L'illicéité de ce type de contenu dépendra de nombreux facteurs, tels que l'origine et l'intention de l'auteur, mais également du contexte dans lequel la publication est réalisée[146] et du sens qui est donné à cette dernière[147]. L'exemple introductif du « *Twitter joke trial* » est symptomatique de la difficulté de modérer le contenu (voir *supra*, I.). En effet, ce genre de déclaration se situe dans une « zone grise » et rend difficile la prise de décision[148]. En outre, l'examen de certains contenus requiert des connaissances juridiques spécifiques, que n'ont pas nécessairement les modérateurs[149].

Le *Digital Services Act* impose aux plateformes numériques, notamment lorsqu'elles décident de supprimer des contenus, d'exposer les faits sur lesquels s'appuie leur décision et de motiver celle-ci, en application du droit ou de leurs conditions générales[150]. La nécessité pour les modérateurs de procéder à un examen factuel et juridique du cas d'espèce est, à notre avis, inconciliable avec la vitesse à laquelle les modérateurs doivent rendre leur décision. En effet, ces derniers reçoivent quotidiennement des millions de signalements, qui doivent être analysés en quelques secondes seulement[151]. Le fait d'imposer aux réseaux sociaux de telles exigences de motivation nous semble particulièrement contraignant. Dans une perspective de sécurité juridique, il revient à notre avis au législateur européen et aux États membres de l'Union de désigner le type de contenus illicites pour lesquels une motivation juridique est exigée, soit en fonction de la gravité de l'infraction, soit en fonction des biens juridiques qui méritent une protection renforcée (par exemple, en matière de droits de propriété intellectuelle).

144 GOLDMAN (n. 4), p. 58 ; YORK/ZUCKERMAN (n. 58), p. 149.

145 CAPLAN (n. 98), p. 25 ; VAN EECKE (n. 13), p. 1465.

146 CAPLAN (n. 98), p. 13 ; K. KLONICK/T. KADRI, *How to Make Facebook's « Supreme Court » Work*, New York Times, 2018, https://perma.cc/T77K-B6K6, consulté le 31.08.2022 ; VAN EECKE (n. 13), p. 1466.

147 Par exemple, un sens humoristique ou politique.

148 LAIDLAW (n. 63), p. 139.

149 VAN EECKE (n. 13), p. 1466. On pense notamment aux atteintes aux droits de propriété intellectuelle.

150 Art. 17 par. 1 et 3 let. *b*, *d* et *e* DSA.

151 R. F. JØRGENSEN, « Rights Talk : In the Kingdom of Online Giants », *in* R. F. JØRGENSEN, *Human Rights in the Age of Platforms*, Cambrige/Londres 2019, p. 163-187, p. 171 ; A. PATI, *Facebook : Moderating 2 Billion. How Moderation Works and Where It Goes*, *Medium*, https://medium.com/dsckiit/facebook-moderating-2-billion-e67f3fbc1c15, consulté le 31.08.2022.

C. Les droits procéduraux des utilisateurs face au pouvoir des réseaux sociaux

L'activité de modération de contenu est susceptible de restreindre les droits fondamentaux des utilisateurs, principalement leur liberté d'expression et leur liberté économique[152]. Une publication licite peut être supprimée car considérée comme contraire aux conditions générales de la plateforme, tandis qu'une publication illicite peut être maintenue à tort. Par ailleurs, lorsque le compte *Instagram* d'un influenceur[153] est supprimé, celui-ci n'est plus en mesure de gagner sa vie. Cette ambiguïté, combinée à la difficulté qu'ont les modérateurs, tant humains qu'automatisés, de déterminer le caractère illicite de certains contenus (voir *supra*, III.B.2.), risque de faire payer un lourd tribut aux libertés fondamentales des utilisateurs.

Par conséquent, il est à notre avis essentiel que les réseaux sociaux – et plus généralement, les plateformes numériques – respectent les droits procéduraux des utilisateurs, en mettant à leur disposition des voies de recours effectives à l'encontre des décisions de modération.

1. *Les voies de recours contre les décisions de modération*

Les réseaux sociaux mettent à disposition de leurs utilisateurs des moyens leur permettant de contester une décision de modération. Sur *Facebook*, une décision de modération peut, en principe, faire l'objet d'une demande d'examen[154]. Dans le cas où la décision est confirmée, les utilisateurs ont la possibilité d'introduire un appel auprès du Conseil de surveillance de *Meta* (*oversight board*), qui est l'autorité suprême des réseaux sociaux *Facebook* et *Instagram*[155].

---

152 BASTIAN (n. 54), p. 53 ; KOLTAY (n. 65), p. 160 : « [T]he moderators [...] gave a far greater influence on the freedom of discussion and exchanges on the platform than does the dedicated government apparatus » ; YORK/ZUCKERMAN (n. 58), p. 137.

153 Un influenceur est une personne qui influence l'opinion, la consommation par son audience sur les réseaux sociaux (Dictionnaire Le Petit Robert, Paris 2022).

154 FACEBOOK, *Mode d'emploi : Je pense que Facebook n'aurait pas dû enlever ma publication*, www.facebook.com/help/2090856331203011?helpref=faq_content, consulté le 31.08.2022 ; KLONICK (n. 122), p. 2434. Un système d'appel est également prévu par *Twitter*, voir TWITTER, *Centre d'assistance : Notre gamme d'options pour l'application de nos politiques*, https://help.twitter.com/fr/rules-and-policies/enforcement-options, consulté le 31.08.2022.

155 FACEBOOK (n. 154).

Lancé en 2019 par *Facebook*, le Conseil de surveillance est une entité extrajudiciaire de résolution des litiges indépendante et composée de juristes, de professeurs d'université et d'anciennes personnalités politiques[156]. Le Conseil de surveillance peut être comparé à une institution arbitrale[157]. En effet, ses décisions sont contraignantes pour les utilisateurs et aboutissent, en principe, soit au rétablissement du contenu supprimé, soit à la confirmation de la suppression[158]. À ce jour, le Conseil de surveillance a rendu 26 décisions, majoritairement dans des affaires à caractère politique[159].

Le Conseil de surveillance sélectionne les appels qu'il souhaite traiter, en fonction de l'importance de l'affaire et de son impact prospectif sur les politiques de modération du groupe *Meta*[160]. L'objectif est de créer des « précédents » qui serviront de base à la régulation de la parole sur Internet[161]. Les cas sélectionnés par le Conseil posent ainsi des questions qui ont trait aux libertés fondamentales des utilisateurs, principalement à leur liberté d'expression et d'opinion.

Il est intéressant de relever que le Conseil de surveillance, une entité privée créée à l'initiative de Mark Zuckerberg, applique un raisonnement similaire à celui d'un juge constitutionnel[162]. En effet, les réseaux sociaux s'engagent, notamment par le biais des Principes directeurs des Nations Unies relatifs aux entreprises et aux droits de l'Homme (UNGP), à respecter et à appliquer les droits fondamentaux dans les relations avec leurs utilisateurs[163]. Dans l'affaire concernant la suspension des comptes *Facebook* et *Instagram* de Donald Trump en raison de l'attaque du Capitol[164], le Conseil de surveillance a appliqué le droit fondamental à la liberté d'expression tel que défini par l'art. 19 du

156 CONSEIL DE SURVEILLANCE, *Le Conseil*, www.oversightboard.com/meet-the-board/, consulté le 31.08.2022. Pour un historique concernant la création du conseil de surveillance, voir KLONICK (n. 122), p. 2448 ss.

157 KLONICK (n. 122), p. 2477.

158 Art. 1 al. 4 par. 1 de la Charte du Conseil de surveillance, https://oversightboard.com/governance/#authority-to-review, consulté le 31.08.2022.

159 CONSEIL DE SURVEILLANCE DE META, *Décisions du Conseil*, www.oversightboard.com/decision/, consulté le 31.08.2022.

160 Art. 2 al. 1 par. 3 de la Charte du Conseil de surveillance (n. 158) ; C. GOANTA/P. ORTOLANI, « Unpacking Content Moderation : The Rise of Social Media Platforms as Online Civil Courts », 2021, p. 16, https://papers.ssrn.com/sol3/papers.cfm?Abstract_id=3969360, consulté le 31.08.2022 (à paraître dans : X. KRAMER *et al.* (édit.), *Frontiers in Civil Justice*, Cheltenham/Northampton 2022).

161 Art. 2 al. 2 par. 2 de la Charte du Conseil de surveillance (n. 158).

162 Voir L. GRADONI, *Constitutional Review via Facebook's Oversight Board : How platform governance had its Marbury v Madison*, Verfassungsblog, 2021, https://verfassungsblog.de/fob-marbury-v-madison/, consulté le 31.08.2022.

163 ONU, *Principes directeurs des Nations Unies relatifs aux entreprises et aux droits de l'Homme*, New York/Genève 2011. Ces principes ne sont toutefois pas contraignants.

164 CONSEIL DE SURVEILLANCE, Décision 2021.001.FB.FBR du 5 mai 2021, https://oversightboard.com/decision/FB-691QAMHJ/, consulté le 31.08.2022.

Pacte ONU II[165]. L'analyse du Conseil a porté sur la légalité, l'intérêt public et la proportionnalité de la décision de suspendre indéfiniment les comptes *Facebook* et *Instagram* de Donald Trump[166]. Dans le cas d'espèce, le Conseil a jugé qu'en maintenant une parole infondée quant à une fraude électorale et un appel persistant à agir, Donald Trump a créé un environnement propice à la violence, en légitimant les émeutes du Capitol. Même si la suspension des comptes était légitimée par un objectif de maintien de l'ordre public, sa durée indéfinie, elle, ne l'était pas. En effet, selon le Conseil, l'absence de limitation de durée de la suspension représentait une restriction disproportionnée à la liberté d'expression de l'utilisateur concerné[167].

La mise à disposition de mécanismes d'appel contre les décisions de modération, tels que ceux mis en place par *Meta*, garantit a priori un semblant d'équité procédurale aux utilisateurs[168]. Toutefois, encore faut-il que ces mécanismes soient effectifs, c'est-à-dire qu'ils respectent certaines garanties procédurales et soient facilement accessibles, rapides et peu onéreux.

2. *Le droit des utilisateurs à un recours effectif*

Sur *Meta,* le premier appel consiste à demander aux modérateurs un (ré)examen de la décision. Cependant, toutes les demandes d'examen ne sont pas examinées, notamment pendant les périodes de forte activité[169]. Le second appel se fait auprès du Conseil de surveillance, lequel sélectionne les cas traités en fonction de leur importance et qui, par conséquent, rend très peu de décisions (voir *supra*, III.C.1.).

Ainsi, même si certains réseaux sociaux mettent à disposition de leurs utilisateurs des mécanismes leur permettant de contester une décision de modération, ces derniers ne semblent pas être effectifs. En effet, dans la grande majorité des cas, les utilisateurs ne sont pas entendus et sont laissés à la merci du pouvoir

165 Pacte international relatif aux droits civils et politiques du 16 décembre 1966 (RS 0.103.2).

166 CONSEIL DE SURVEILLANCE, Décision 2021.001.FB.FBR (n. 164).

167 *Ibid.*

168 KLONICK (n. 122) ; *contra* : BLOCH-WEHBA (n. 94), p. 93, qui estiment que la procédure mise en place par le Conseil de surveillance n'offre qu'un « simulacre de procès équitable ».

169 FACEBOOK (n. 154) : « Pendant les périodes d'activité intense, il ne nous est pas possible d'examiner toutes les demandes [...]. Si votre contenu a été retiré et que vous estimez qu'il ne va pas à l'encontre de nos Standards de la communauté, vous pouvez demander à ce qu'il soit examiné à nouveau et nous réétudierons notre décision si nous le pouvons [...]. Il est impossible de demander un nouvel examen pour certains types de contenu ».

décisionnel des modérateurs[170]. De plus, on peut légitimement se demander si les modérateurs qui procèdent au réexamen des décisions de modération, au stade du premier appel, sont réellement indépendants et impartiaux.

Il est à notre avis essentiel d'établir un système de modération qui, d'une part, respecte des garanties procédurales minimales et, d'autre part, offre aux utilisateurs un accès à une forme de justice rapide, transparente et peu onéreuse[171]. À défaut, les droits fondamentaux des utilisateurs se verraient vidés de leur substance face au pouvoir décisionnel des plateformes numériques, ce qui conduirait à un déni de justice[172]. En ce sens, le Rapporteur spécial des Nations Unies sur la liberté d'expression a identifié l'obligation de fournir un recours effectif comme l'un des aspects les plus importants en ce qui concerne les activités de modération réalisées par les entreprises privées[173].

Définir précisément les principes procéduraux applicables en matière de modération de contenu sortirait du cadre de la présente contribution. Par conséquent, nous nous limiterons à énumérer, sans prétention aucune à l'exhaustivité, quelques principes procéduraux qui s'appliquent – ou devraient s'appliquer – aux procédures de modération de contenu.

Au préalable, les mécanismes de modération peuvent être analysés sous l'angle du cadre légal développé en matière de résolution alternative des litiges (*alternative dispute resolution*, ADR) et de résolution des litiges en ligne (*online dispute resolution*, ODR)[174]. En droit européen, la Directive 2013/11/UE relative au règlement extrajudiciaire des litiges de consommation[175] et le Règle-

170 HATEAID (n. 124) ; KOLTAY (n. 65), p. 182 ; M. SCHEFER/R. CUENI, *Öffentlichkeit Im Wandel : Überlegungen Aus Grundrechtlicher Sicht*, Rapport pour l'OFCOM, 2020, p. 56, https://docplayer.org/199529362-Oeffentlichkeit-im-wandel-ueberlegungen-aus-grundrechtlicher-sicht.html, consulté le 31.08.2022.

171 APPELMAN *et al.* (n. 97), p. 13 ; H. FELD, *Case for the Digital Platform Act : Market Structure and Regulation of Digital Platforms*, New York 2019, p. 29 ; R. VAN LOO, « Federal Rules of Platform Procedure », *The University of Chicago Law Review*, 2021, p. 829-895, p. 867.

172 APPELMAN *et al.* (n. 97), p. 1 ; GOANTA/ORTOLANI (n. 160), p. 8.

173 CONSEIL DES DROITS DE L'HOMME, *Rapport du Rapporteur spécial sur la promotion et la protection du droit à la liberté d'opinion et d'expression*, 6 avril 2018 (A/HRC/38/35), par. 11, 38 et 58.

174 GOANTA/ORTOLANI (n. 160), p. 7, qui soulignent l'intérêt d'examiner la modération de contenu à travers le prisme des ADR ; P. ORTOLANI, « Digital Dispute Resolution : Blurring the Boundaries of ADR », *in* L.A. DIMATTEO *et al.* (édit.), *The Cambridge Handbook of Lawyering in the Digital Age*, Cambridge 2021, p. 140-158, p. 148 ss, qui évoque les « plateformes en tant que fournisseurs de services de résolution des litiges ».

175 Directive 2013/11/UE du Parlement européen et du Conseil du 21 mai 2013 relative au règlement extrajudiciaire des litiges de consommation (Directive ADR ; JO L 165/63).

ment 524/2013 du 21 mai 2013 relatif au règlement en ligne des litiges de consommation[176] fixent les exigences procédurales que doivent respecter les entités de règlement extrajudiciaire des litiges entre consommateurs et professionnels établies dans l'Union européenne[177]. La CNUDCI a également émis, en 2017, des « Notes techniques » non contraignantes relatives aux ODR utilisés dans le cadre du e-commerce[178]. Les ADR et ODR se doivent ainsi de respecter les exigences d'indépendance et d'impartialité[179], d'être accessibles, rapides et peu onéreux[180] et de garantir le droit des parties à un procès équitable[181].

En 2018, plusieurs organisations privées ont adopté les Principes de Santa Clara relatifs à la transparence et la responsabilité en matière de modération de contenu[182]. Ces principes établissent des standards procéduraux que les intermédiaires en ligne doivent prendre en compte lors de leurs activités de modération[183]. Des systèmes de notification et des voies de recours effectives doivent notamment être introduits[184]. En outre, les modérateurs doivent être compétents et comprendre la langue, la culture et le contexte politico-social des publications qu'ils sont amenés à modérer[185].

Le Règlement sur les services numériques (DSA) fixe des obligations précises en matière de modération de contenu (voir *supra*, II.A.3.). Les plateformes numériques devront introduire des mécanismes permettant aux utilisateurs de contester les décisions de modération de contenu, par le biais d'un système interne de traitement des réclamations[186]. Les décisions de modération devront

176 Règlement 524/2013 du Parlement européen et du Conseil du 21 mai 2013 relatif au règlement en ligne des litiges de consommation (Règlement ODR ; JO L 165/1).

177 La Directive ADR et le Règlement ODR sont deux instruments législatifs liés et complémentaires (Directive ADR, consid. 12). Toutefois, ces textes ne s'appliquent pas aux procédures de modération gérées par un réseau social (voir art. 2 par. 2 let. *b* de la Directive ADR).

178 CNUDCI, *Notes techniques sur le règlement des litiges en ligne*, New York 2017. Même si les notes techniques ont pour champ d'application le e-commerce, celles-ci sont à notre avis utiles pour fixer un cadre procédural en matière de modération de contenu.

179 Art. 6 de la Directive ADR ; CNUDCI (n. 178), par. 13 s., p. 3.

180 Art. 8 de la Directive ADR ; CNUDCI (n. 178), par. 9, p. 2.

181 Art. 9 de la Directive ADR ; CNUDCI (n. 178), par. 7, p. 2 et par. 49*c*, p. 9.

182 The Santa Clara Principles on Transparency and Accountability in Content Moderation, version 2.0, 2021, https://santaclaraprinciples.org/, consulté le 31.08.2022. D'autres textes non contraignants ont été adoptés par des organisations privées (voir GOLDMAN (n. 4), p. 14 ss).

183 Art. 1 des Principes de Santa Clara (n. 182) : « Companies should ensure that human rights and due process considerations are integrated at all stages of the content moderation process [...] ».

184 Art. 5 des Principes de Santa Clara (n. 182).

185 Art. 3 des Principes de Santa Clara (n. 182).

186 Art. 20 DSA ; consid. 58 DSA.

être précisément motivées, tant factuellement que juridiquement[187]. Le Règlement introduira également un mécanisme de certification des entités extrajudiciaires de résolution des litiges (ADR)[188]. Afin d'être certifiées au sens du Règlement, ces entités devront offrir aux utilisateurs des garanties procédurales suffisantes, notamment être indépendantes et impartiales, disposer de l'expertise nécessaire pour analyser l'illicéité d'un contenu et offrir une procédure rapide et peu onéreuse[189].

Finalement, l'utilisation de mécanismes automatisés de modération (voir *supra*, III.A.) doit faire l'objet d'une transparence renforcée (*algorithmic accountability*)[190]. Lorsque des moyens automatisés sont utilisés par la plateforme pour repérer, filtrer et supprimer du contenu, les utilisateurs doivent être informés de l'automatisation à la base de la décision, afin de pouvoir, cas échant, demander un réexamen de la décision par un être humain. Ce « droit à une décision humaine »[191] a d'ailleurs été consacré par le Conseil de surveillance de *Meta*, en application de l'art. 2 du Pacte ONU II relatif au droit à un recours effectif[192]. Selon le Conseil, lorsque *Facebook* recourt à des moyens automatisés pour modérer le contenu, la plateforme doit assurer à ses utilisateurs une possibilité de faire appel de la décision auprès d'un être humain[193].

187 Art. 17 et 20 par. 5 DSA. Les modérateurs devront préciser les faits sur lesquels s'appuie la décision et, lorsque la décision concerne des contenus prétendument illicites, donner une référence au fondement juridique sous-jacent et des explications des motifs pour lesquels ces informations sont considérées comme des contenus illicites (art. 17 par. 3 let. *b* et *d* DSA).

188 Art. 21 DSA.

189 Art. 21 par. 3 DSA.

190 Art. 15 par. 1 let. *c* et *e*, 16 par. 6, et 17 par. 3 let. *c* DSA.

191 HUQ (n. 123), p. 615 ; D. SANCHO, « Automated Decision-Making under Article 22 GDPR », *in* M. EBERS/S. NAVAS (édit.), *Algorithms and Law*, Cambridge 2020, p. 136-156, p. 147 ss.

192 CONSEIL DE SURVEILLANCE, Décision 2020-004-IG-UA du 28 janvier 2021, https://oversightboard.com/news/682162975787757-oversight-board-overturns-original-facebook-decision-case-2020-004-ig-ua/, consulté le 31.08.2022. Cette affaire concerne la suppression, par des moyens automatisés, d'une image publiée sur *Instagram* qui avait pour objectif de sensibiliser la population au dépistage du cancer du sein.

193 CONSEIL DE SURVEILLANCE, Décision 2020-004-IG-UA (n. 192) ; voir ég. l'art. 20 par. 6 DSA, selon lequel « [l]es fournisseurs de plateformes en ligne veillent à ce que les décisions [rendues sur réclamation] soient prises sous le contrôle de collaborateurs dûment qualifiés, et pas uniquement par des moyens automatisés ».

3. *L'obligation de l'État de garantir les droits procéduraux des utilisateurs face aux réseaux sociaux*

L'État a une obligation positive de protéger les particuliers contre les atteintes à leurs libertés fondamentales par d'autres particuliers, notamment par le biais du droit positif[194]. Cette obligation est accentuée lorsque les libertés fondamentales d'un grand nombre de personnes sont menacées par des entités privées socialement puissantes, telles que les grandes plateformes numériques[195]. En effet, le fait que des acteurs privés, plutôt que des tribunaux étatiques, décident de la légalité de certains contenus soulève des préoccupations tant du point de vue de la liberté d'expression, que du droit à un procès équitable (*due process*)[196].

Par conséquent, il revient au législateur d'adopter des mesures permettant de protéger les utilisateurs dans le cadre d'une procédure de modération de contenu[197]. En adoptant le Règlement sur les services numériques, le législateur européen prend les mesures visant à renforcer les droits procéduraux des utilisateurs dans leur relation de dépendance vis-à-vis des plateformes numériques[198].

Le législateur américain est également en passe de modifier son régime de responsabilité indirecte, en imposant des obligations plus strictes en matière de modération de contenu. Deux projets de loi ont pour objectif de réduire l'étendue de l'immunité accordée aux plateformes numériques en vertu de la section 230 CDA (voir *supra*, II.B.)[199].

IV. Conclusion

Le présent ouvrage a pour titre « la technologie, l'humain et le droit ». Afin de clore nos propos, nous proposons donc de reprendre ces trois éléments qui, à notre avis, sont centraux dans le cadre de la modération de contenu.

194 CourEDH, *Mosley c. le Royaume-Unis* du 10 mai 2011 (requête n° 48009/08), par. 106 ; CONSEIL DES DROITS DE L'HOMME (n. 173), par. 6 ; CALLAMARD (n. 87), p. 199 et 201 ; P. MAHON, *Droit constitutionnel – Droits fondamentaux*, 3e éd., vol. II, Bâle/Neuchâtel, N 28, p. 49.

195 CONSEIL DES DROITS DE L'HOMME (n. 173), par. 6 ; Rapport OFCOM (n. 6), p. 31.

196 JORGENSEN (n. 151), p. 176.

197 MACCARTHY (n. 66), p. 116 ; VAN LOO (n. 171), p. 875.

198 COMMISSION EUROPÉENNE (n. 33), p. 14.

199 CONGRÈS DES ÉTATS-UNIS D'AMÉRIQUE, *Projets de Digital Platform Commission Act of 2022* du 12 mai 2022 (S.4201) et de *21st Century FREE Speech Act* du 27 avril 2021 (S. 1384).

La *technologie* façonne la manière dont se déroule l'échange en ligne d'information et d'opinion. Le code (*i.e.* les algorithmes) dicte le *modus operandi* de la plateforme, dont l'objectif principal est de capitaliser sur les données personnelles de ses utilisateurs. Les algorithmes des réseaux sociaux sélectionnent ce que les utilisateurs reçoivent dans leur fil d'actualité, en fonction de leurs interactions précédentes sur la plateforme. De plus, des moyens automatisés sont utilisés à des fins de modération de contenu. L'usage croissant de la technologie à des fins de modération de contenu permet à la plateforme d'influencer l'opinion des *utilisateurs*. Ces derniers se trouvent à la merci du pouvoir décisionnel de la plateforme, de ses algorithmes et des modérateurs humains employés par celle-ci.

Par conséquent, on peut légitimement se demander quel est le rôle du *droit* dans le cadre de la modération de contenu. Il revient à notre avis au législateur la tâche – certes ardue mais essentielle – de trouver un juste équilibre entre, d'une part, la protection des justiciables sur Internet et, d'autre part, l'objectif de ne pas freiner le progrès technologique en fixant des exigences légales trop contraignantes aux plateformes. À l'instar du Règlement sur les services numériques, cet équilibre cherchera à fixer des standards minimaux applicables à une procédure de modération de contenu, tout en assurant que les utilisateurs aient accès à une forme de justice efficiente, rapide et peu onéreuse : une justice numérique.