# Modération

**Source**
*reforme.net*
Gilles Dowek
4 septembre 2019

## 1. Liberté d'expression : le dilemme des réseaux sociaux

*Le Web et les réseaux sociaux offrent une liberté de parole qui ne va pas sans dérives. Comment les réguler quand les valeurs et les codes diffèrent selon les pays ?*

Les questions posées par les tentatives de réguler le Web et les réseaux sociaux actualisent un vieux dilemme éthique : nous chérissons le fait de nous exprimer librement, ainsi que celui de vivre en paix, sans être injuriés, harcelés, menacés ou manipulés. Que faire quand ces valeurs entrent en conflit ?

Faut-il interdire les propos injurieux, limitant ainsi la liberté d'expression ? Ou faut-il les autoriser, limitant alors la possibilité de vivre en paix ? Face à un tel dilemme, notre éthique et notre droit proposent de fragiles compromis, interdisant, par exemple, les menaces de mort, mais autorisant la dérision. Ces compromis sont toujours locaux et injustes – certaines injures se révélant mieux tolérées que d'autres. Ainsi, ceux trouvés en Europe valorisent davantage la possibilité de vivre en paix ; et en Amérique du Nord, la liberté d'expression.

**Anonymat ou pas ?**

Dans le cas du Web et des réseaux sociaux, ce dilemme se double d'un second : faut-il autoriser, ou non, les utilisatrices et les utilisateurs de ces médiums à porter un masque, comme les Vénitiens pendant le carnaval ?

Autoriser les masques revient à abandonner tout espoir de poursuivre et de condamner les auteurs de propos injurieux. En revanche, interdire ces masques tend à instaurer un régime panoptique où un Big Brother hyper-mnésique observe en permanence les propos tenus par chacun, nécessairement sous son état civil complet. Mais rassurez-vous : si vous n'avez rien à cacher, vous n'avez rien à craindre.

Cette situation est rendue plus complexe encore, dans le cas du Web et des réseaux sociaux, par deux éléments. Le premier est que les compromis trouvés sont souvent locaux, quand les réseaux eux-mêmes sont mondiaux et concernent des êtres aux sensibilités différentes. Aussi, cela nous oblige-t-il à chercher des compromis avec des personnes que la nudité de la Vénus de Botticelli ou du David de Michel-Ange peut choquer davantage qu'un appel au meurtre, par exemple.

Le second est que le Web et les réseaux sociaux ont donné à toutes et à tous un accès à la parole publique, naguère réservée à peu, multipliant l'ampleur de ces dilemmes. Il est ainsi possible que le nombre de propos injurieux n'ait pas augmenté récemment, mais que le changement réel soit qu'ils sont désormais tenus, non au comptoir d'un bistrot devant trois autres individus, mais devant sept milliards d'humains. Or, cet accès universel à la parole publique – qui en soi est une bonne chose – permet, hélas, à la foule de se substituer au juge.

Ainsi, il est inquiétant qu'une journaliste écrive : « Toi aussi raconte, en donnant le nom et les détails, le harcèlement sexuel que tu as connu dans ton boulot. » Mais il est également grave qu'un tel message soit perçu comme positif par des millions d'hommes et de femmes, comme si la gravité du délit dispensait d'un procès équitable et autorisait la foule à accuser sans preuve.

**Le droit des victimes**

Que faire alors face à ces dilemmes ? C'est ici que la distinction entre « juger » et « sanctionner » peut contribuer à construire un compromis acceptable. Un jugement sert avant tout à reconnaître qu'une victime a subi un préjudice et qu'un auteur en est responsable. Il doit aussi apporter une réparation à la victime. La sanction du responsable, en revanche, est presque secondaire. Et il faut admettre qu'elle est même impossible dans certains cas extrêmes, par exemple, quand cela concerne la sortie d'un conflit armé. Mais cela ne supprime nullement le droit des victimes à la vérité, à la justice et à la réparation.

Transposé dans le cadre, heureusement moins tragique, du Web et des réseaux sociaux, cela signifie que la régulation des messages injurieux doit avoir pour but premier la protection des victimes.

La sanction des responsables, qui est rarement possible, se retrouve, dans ce cas également, secondaire. Et il n'est pas nécessaire d'exiger des utilisatrices, des utilisateurs du Web et des réseaux sociaux une transparence inquisitrice, qui restreindrait de fait leur liberté d'expression, pour protéger les victimes.

**Responsabilité individuelle**

En revanche, cette protection impose de supprimer immédiatement un message injurieux – et toutes ses répliques sur le même réseau, voire sur d'autres. Comme toute mesure de réparation, elle est nécessairement partielle, mais elle contribue à reconnaître le préjudice subi par la victime et rappelle que l'injure est inacceptable.

Cette responsabilité doit naturellement être imposée par la loi aux entreprises qui administrent ces plateformes. Mais elle est aussi la nôtre car elle se situe à un niveau individuel. En effet, nous devons prendre conscience qu'en répliquant de tels messages, nous en sommes nous aussi responsables.

**Source**
*lemonde.fr*
Perrine Signoret
18 avril 2017

## 2. Une vidéo de meurtre sur Facebook, dernier dérapage d'une longue série

*Il a fallu attendre deux heures avant que le réseau social ne reçoive le premier signalement d'internaute. Le compte du meurtrier a été désactivé une vingtaine de minutes plus tard.*

Vers 11 heures, dimanche 16 avril, un Américain de 37 ans, du nom de Steve Stephens, a commencé à poster sur sa page Facebook une macabre série de vidéos. Dans la première d'entre elles, il proclame son envie de tuer quelqu'un. Dans la deuxième, il met ses menaces à exécution, assassinant face caméra un homme de 74 ans. Dans la troisième, il reconnaît, en direct cette fois-ci, le meurtre, ainsi que treize autres (pour l'heure non confirmés par les autorités). La police a annoncé, mardi 18 avril, que Steve Stephens s'était ensuite donné la mort.

Il aura fallu attendre deux heures après la première publication de l'une de ces vidéos pour qu'un internaute la signale au réseau social, grâce au bouton prévu à cet effet, situé dans une liste déroulante au-dessus de chaque post. Facebook est ensuite intervenu rapidement : vingt-trois minutes plus tard, le compte du suspect était désactivé.

S'il est difficile d'obtenir du réseau social un chiffre précis, la plupart des signalements seraient traités en moins de soixante-douze heures. Certains sont prioritaires : contenus s'apparentant à du harcèlement, à des menaces physiques ou à des faits de violence. Mais un délai de quelques heures est généralement nécessaire avant que les centaines de modérateurs de Facebook ne se saisissent de ces cas.

**« Nous travaillons dur pour garder un environnement sécurisé »**

Le lendemain des faits, les responsables du réseau social ont d'ailleurs souligné, dans un communiqué, la relative rapidité de leur action, tout en assurant vouloir « faire mieux » : « C'est un crime horrible, nous n'autorisons pas ce type de contenus. Nous travaillons dur pour garder un environnement sécurisé et restons en contact avec les autorités et les services de secours lorsque des menaces directes à l'intégrité physique surviennent. »

Cette déclaration a été assortie de la promesse de quelques améliorations, qui concernent principalement le système de signalement de vidéos, photographies ou autres violant les règles

d'utilisation du réseau social.

Pas question, toutefois, de modifier le fonctionnement intrinsèque du réseau social – par exemple, le fait qu'un contenu ne soit vérifié par les équipes de modération que s'il est signalé par des utilisateurs. Il s'agit de le rendre « aussi rapide et facile que possible », a souligné Facebook, sans toutefois préciser comment. Un changement d'autant plus nécessaire que l'assassinat de dimanche n'est pas le premier crime à être filmé et mis en ligne sur Facebook.

Depuis son ouverture à tous les utilisateurs en février 2016, l'outil Facebook Live, qui permet de diffuser des vidéos en direct, a été à plusieurs reprises le théâtre de ce phénomène sordide. Il y a deux semaines, le viol collectif présumé d'une mineure américaine était diffusé ; quelques mois plus tôt, une autre publication montrait un homme séquestré, insulté et frappé par quatre agresseurs, dans la banlieue de Chicago.

**Malaise en direct**

Au total, une cinquantaine d'événements violents, dont plusieurs meurtres et suicides, auraient été diffusés via Facebook Live, selon le décompte du Wall Street Journal.

Certains ont particulièrement marqué les esprits, comme la mort en 2016, dans l'Arkansas, de Keiana Herndon, une mère de famille âgée de 25 ans. Atteinte d'un cancer, elle avait pris l'habitude de filmer son quotidien pour ses amis. Un jour, en plein direct sur Facebook, elle fait un malaise ; lorsque l'un de ses amis arrive pour tenter de la sauver, trente minutes plus tard, il ne peut que constater son décès. Parmi les « spectateurs » de son « live » (pourtant de plus en plus nombreux), aucun n'aurait, selon la famille de la jeune femme, appelé les secours.

Autre plate-forme, même constat : le 10 mai 2016, sur Periscope, une application pour smartphone de partage de vidéos, une Française de 19 ans s'est donné la mort devant un millier d'internautes. La jeune fille s'est jetée sous une rame du RER dans l'Essonne, en région parisienne, après avoir prévenu que quelque chose de « très choquant » allait se passer.

Si les politiques de modération de Facebook ou d'autres réseaux sociaux font régulièrement débat, le comportement voyeuriste des internautes interroge. Selon le New York Times, ils seraient plusieurs « millions » à avoir ainsi visionné l'une des trois vidéos de Steve Stephens.

Certains les ont aussi téléchargées, avant d'en diffuser des extraits sur Facebook et Twitter… Une pratique autorisée par Facebook, à condition que cela serve à des fins d'information, ou à dénoncer des comportements dangereux.

Est-ce le passage à la postérité qui incite des criminels à partager ces images sur les réseaux sociaux ? Cette recherche de célébrité ne date pas, en tout cas, de l'avènement de ces derniers. On l'appelle ainsi fréquemment « syndrome d'Erostrate », en référence à cet homme qui incendia le temple d'Artémis à Ephèse dans le seul but d'être connu, un épisode remontant à... l'antiquité grecque.

**Source**
*lefigaro.fr*
Lucie Ronfaut
19 mars 2019

## 3. Christchurch : pourquoi la modération des réseaux sociaux est-elle inefficace ?

*Les grandes plateformes en ligne, surtout YouTube (propriété de Google) et Facebook, ont eu du mal à empêcher la diffusion de vidéos portant sur l'attentat contre deux mosquées en Nouvelle-Zélande, le 15 mars 2019.*

L'attentat de Christchurch, qui a fait 50 morts, a démontré encore une fois les faiblesses de la modération des géants du Web. Le terroriste a diffusé son massacre en direct sur Facebook, dans une vidéo longue de 17 minutes. Le document s'est ensuite propagé sur de nombreux sites, notamment YouTube (propriété de Google), Instagram (propriété de Facebook) et Twitter, dans son intégralité ou via des extraits. De nombreux internautes se sont émus de cette large diffusion et se sont interrogés, à juste titre, sur l'efficacité des outils de modérations des réseaux sociaux.

Ce week-end, Facebook a donné un peu plus de détails sur ses propres efforts pour endiguer la diffusion de la vidéo. « La police néo-zélandaise nous a avertis de l'existence de ce live. Nous avons vite supprimé les comptes Facebook et Instagram du tireur », a expliqué Mia Garlick, responsable de Facebook en Nouvelle-Zélande, via le compte Twitter officiel du réseau social. « Dans les premières 24 heures [de la diffusion de la vidéo], nous avons supprimé 1,5 million de vidéos de l'attaque dans le monde, dont 1,2 million ont été bloquées avant même d'être publiées. » Le réseau social a également précisé que les versions modifiées de la vidéo, même celles qui ne montraient pas de contenus violents, avaient aussi été supprimées, par respect pour les victimes. Sur YouTube, des copies de la vidéo de l'attentat étaient publiées toutes les secondes, au plus fort de la crise, forçant la plateforme à désactiver temporairement certains termes de recherche liés à l'évènement, afin de limiter leur visibilité.

Ces détails confirment toute la complexité que représente l'exercice de contrôle des contenus sur Internet. Et démontrent que la modération des réseaux sociaux souffre encore de nombreuses failles, certaines difficilement dépassables.

**L'intelligence artificielle, une solution imparfaite**

Dans certains cas bien précis, les géants du Web recourent à des outils d'intelligence artificielle afin d'empêcher la publication de contenus problématiques, ou de les repérer et les signaler automatiquement à des modérateurs humains. C'est notamment le cas des photos ou vidéos pornographiques, pédopornographiques ou de contenus de propagande terroriste. Dans ces deux dernières catégories, les plus grands acteurs du secteur ont recours à des banques de données communes. Elles leur permettent de partager les photos et les vidéos déjà repérées, afin que chacun puisse en empêcher la publication future sur sa propre plateforme. Il s'agit d'un système dit d' « empreinte numérique ». Mais ce type d'outil ne peut être utilisé que dans le cas de contenus déjà connus. Pour l'attentat de Christchurch, il n'aurait pas pu empêcher la toute première diffusion de la vidéo. Par ailleurs, les internautes savent aussi tromper les machines, en modifiant légèrement les contenus copiés, échappant ainsi à cette reconnaissance automatique.

Si l'intelligence artificielle est beaucoup utilisée pour les contenus de pédopornographies ou de propagande terroriste, elle l'est beaucoup moins pour d'autres catégories. Par exemple, la moitié des contenus d'incitation à la haine supprimés de Facebook sont encore signalés par des internautes, plutôt que d'être repérés par des algorithmes. Pour le harcèlement, cette proportion atteint 85 %. Une machine n'est pas encore capable de comprendre la subtilité du langage humain et ses intentions lorsqu'il s'agit d'insultes ou de menaces.

**Des modérateurs humains pas assez nombreux**

Google, Facebook ou Twitter appliquent tous une modération dites *a posteriori.* La plupart du temps, il est nécessaire qu'un internaute signale un contenu problématique pour qu'il soit ensuite contrôlé par un modérateur humain. On connaît peu de choses sur les personnes qui sont employées par les grandes plateformes pour modérer leurs contenus. Dans le cas de Facebook, on sait qu'elles sont environ 15'000 dans le monde, en grande majorité salariées d'entreprises sous-traitantes. Régulièrement, des articles, reportages et enquêtes témoignent des conditions de travail difficile de ces modérateurs, submergés de travail et exposés à des contenus violents à longueur de journée.

Dans le cas de l'attentat de Christchurch, le premier signalement d'un internaute est intervenu 29 minutes après la publication de la vidéo, et 12 minutes après la fin du live. Le contenu a été vu au total 4000 fois avant d'être supprimé par Facebook.

**Des règles complexes et changeantes**

Les modérateurs humains effectuent leurs décisions selon les lois en vigueur dans le pays d'origine du contenu (par exemple, en France, il est interdit de publier des contenus niant l'existence de l'Holocauste, ce qui n'est pas le cas aux États-Unis) et les règles internes du réseau social en question. Ces dernières sont complexes, et évoluent régulièrement. Par exemple, dans le cas de Facebook, il n'est pas autorisé de publier un contenu « qui glorifie la violence, ou qui prône la

souffrance ou l'humiliation d'autres personnes ». Il est en revanche autorisé de publier certains contenus explicites dans des fins d'information ou de dénonciation, « afin d'aider les personnes à sensibiliser leur communauté à différents problèmes ». Ces règles peuvent être outrepassées au cas par cas. Dans le cas de l'attentat de Christchurch, Facebook a fait le choix de supprimer toutes les vidéos reprenant les images de l'attaque, quelles que soient les intentions des internautes.

**L'urgence du « live »**

Dans tous les cas, ni l'intelligence artificielle, ni les modérateurs humains ne sont totalement efficaces à gérer un évènement qui se déroule en temps réel, surtout lorsqu'il s'agit d'une vidéo live, difficile à repérer en pleine diffusion. Dans les situations d'urgence, les grandes plateformes optent généralement pour un mélange des deux. Par exemple, YouTube a fait le choix d'empêcher automatiquement la republication de la vidéo de l'attentat de Christchurch en entier, mais d'envoyer celles reprenant seulement des extraits à ses modérateurs humains, afin de jauger s'il s'agissait de contenus publiés à des fins d'information.

**Des plateformes immenses et conçues pour la viralité**

Reste un dernier élément compliquant ce travail de modération, sans doute le plus difficile à régler : la nature même des géants du Web. Facebook, Instagram ou YouTube sont d'immenses plateformes, fréquentées des milliards de personnes dans le monde, publiant chaque jour un nombre phénoménal de contenus. Ainsi, d'après les déclarations de Facebook, le réseau social est parvenu à bloquer 80 % des vidéos montrant la tuerie de Christchurch avant même leur publication. Il s'agit d'un ratio plutôt encourageant. Néanmoins, parce que Facebook est une si grande plateforme, cela signifie que 300'000 vidéos liées à l'attentat, les 20 % restants, ont tout de même pu être publiées. Ce chiffre est conséquent. D'autant que ces contenus ont bénéficié d'une forte viralité, poussés par des algorithmes de recommandation entraînés à mettre en avant les publications provoquant beaucoup de réactions. Plus que de moyens humains, plus que de nouveaux outils technologiques, c'est aussi d'une révolution interne dont auraient besoin les réseaux sociaux pour mieux lutter contre la violence en ligne.

**Source**
*The Conversation*
Christine Dugoin-Clément
30 mai 2019

# 4. Christchurch.0 : quel espoir pour l'Internet de demain ?

Le 15 mars 2019, Brenton Tarrant, connu pour son extrémisme de droite, attaquait deux mosquées dans la ville de Christchurch, en Nouvelle-Zélande, pendant la prière du vendredi. Un bilan tragique : 51 morts et de 49 blessés.

Si l'ampleur de cet acte a soulevé une vague de choc dans ce pays, considéré comme le second plus sûr au monde, un deuxième traumatisme arrivait avec la diffusion en « live », sur Facebook, de l'agression meurtrière.

17 minutes d'images, avant que leur émission ne soit interrompue. Alors que l'hébergeur affirma avoir rapidement supprimé les comptes Facebook et Instagram du meurtrier après l'avertissement lancé par la police néo-zélandaise, Facebook déclarait avoir supprimé un total de 1,5 million de vidéos de l'attaque en 24 heures, dont 1,2 million bloquées au téléchargement.

**Le script : l' appel de Christchurch et Paris**

La problématique des plates-formes d'intermédiation dans la diffusion de contenus haineux a déjà été soulignée à plusieurs reprises. En France, le président Macron a annoncé la mise en œuvre d'une loi contre les contenus haineux en ligne, en février 2019, lors du dîner au Conseil représentatif des institutions juives (CRIF). Le 11 mars 2018, la députée Laetitia Avia (LREM), en coopération avec l'écrivain Karim Amellal et le vice-président du CRIF Gil Taieb, a fait suivre à cette déclaration une proposition de loi devant l'Assemblée nationale.

Mais c'est aujourd'hui « l'appel de Christchurch » pour lutter contre le terrorisme et l'extrémisme

violent en ligne qui a résonné dans tous les médias. L'initiative lancée par la première ministre néo-zélandaise Jacinda Ardern et par le Président français Macron à Paris le 15 mai 2019 vise à construire un front globalisé contre la haine et l'incitation à la violence sur Internet.

Le sommet survient après un autre appel, celui de Paris, lancé le 12 novembre 2018, lors de la réunion du Forum de gouvernance de l'Internet (FGI) à l'Unesco. Il s'agit d'une incitation à rétablir la confiance et la sécurité dans le cyberespace.

Bien que ces appels – preuves d'une volonté politique forte – soient indispensables et vertueux, il faut s'interroger sur leur capacité à être mis en œuvre, à devenir universels et sur leur efficacité. Seules la coopération des plates-formes et l'implication des internautes assureront l'efficacité recherchée, le tout sous l'égide d'États qui devront faire respecter leurs décisions par les géants du numérique, sans aller trop loin dans l'étendue des contrôles mis en place.

**L'acteur 1 : la volonté politique**

La volonté politique est un préalable à toute lutte contre les contenus haineux et les incitations à la violence en ligne. À ce titre, l'appel de Christchurch, qui ne vise aucun mouvement terroriste en particulier, est emblématique. Ce texte est rédigé de façon à pouvoir être adopté par le plus grand nombre d'acteurs, qu'ils soient publics ou privés. Cependant, il reste assez généraliste et ne propose pas de mesures concrètes, se bornant à renvoyer à des discussions ultérieures, notamment lors des G7 et G20, pour l'adoption de décisions effectives.

D'où les premiers écueils : la concrétisation de tels appels impliquent qu'ils soient adoptés par tous ou, a minima, par une large majorité. En cas de contravention, cela empêcherait les États de se retrancher derrière des juridictions différentes ou plus favorables. De fait, immédiatement après l'appel de Christchurch, l'Australie, le Canada, la Commission européenne, la France, l'Allemagne, l'Indonésie, l'Inde, l'Irlande, l'Italie, le Japon, la Jordanie, les Pays-Bas, ont adopté le texte, alors que les États-Unis annonçaient qu'ils ne le ratifieraient pas. La Maison Blanche indiquait, néanmoins, qu'elle continuerait d'appuyer les objectifs généraux présentés dans cet appel en déclarant :

> « Nous encourageons les entreprises technologiques à appliquer leurs conditions de service et leurs normes communautaires interdisant l'utilisation de leurs plates-formes à des fins terroristes. Nous continuons d'être proactifs dans nos efforts pour lutter contre le contenu terroriste en ligne, tout en continuant de respecter la liberté d'expression et la liberté de la presse. »

**Le figurant : la Maison Blanche**

Sous couvert notamment du respect de la liberté d'expression, visée par le premier amendement, l'un des principaux acteurs sur la scène internationale s'est désolidarisé de l'appel. Déjà, certains analystes pointent du doigt que l'attachement à la liberté d'expression, mis en avant par l'administration de Trump, permet à de nombreux républicains de publier des tweets qui seraient probablement interdits sous d'autres juridictions.

Néanmoins, la distance prise par les États-Unis fragilise le texte et peut avoir des conséquences désastreuses. Déjà, certains médias dont Russia Today (RT), ont rebondi sur cette prise de position, expliquant que le président américain avait probablement sauvé la liberté d'expression, et accusant en filigrane les États signataires de vouloir porter atteinte (voir annihiler) la liberté d'expression.

**Des risques de revers**

Tout d'abord, il apparaît que les petites plates-formes n'ont pas été associées aux travaux préparatoires. Si cette omission résulte davantage de l'urgence que d'une volonté d'exclusion, les conséquences pourraient néanmoins être les mêmes. Concrètement, ces dernières pourraient se trouver dans l'impossibilité pour de s'aligner sur les prérequis exigés.

De plus, des divergences de régulation entre les États peuvent conduire à de différents Internet ou, plus exactement, à un accès différencié aux contenus et à l'information. Selon leur pays d'origine, les internautes pourraient ainsi ne pas accéder aux mêmes contenus et services que ceux du pays voisin, comme le fait remarquer l'Internet Society.

Enfin, des régulations trop différenciées fragmenteraient l'Internet : une menace à la coexistence de la lutte contre la haine avec le maintien du droit à l'information.

**L'acteur 2 : Internet**

Avant même l'appel de Christchurch et en réponse à l'horreur suscitée par la diffusion en « live » de la tuerie, Facebook annonçait réduire la possibilité de diffuser des vidéos en direct sur sa plate-forme. Toute personne ne respectant pas la politique de diffusion du réseau social serait ainsi impactée par cette restriction. La sanction serait alors une interdiction d'utilisation du service « live » pour une durée de 30 jours.

La société de Zuckerberg annonçait également vouloir limiter les copies des vidéos violentes, un autre aspect du problème. Ce qui est le cas pour la vidéo de Tarrant : même si la vidéo a été retirée, elle a pu être téléchargée avant sa suppression, et peut continuer ainsi à circuler.

Facebook n'est cependant pas la seule société inquiétée dans la diffusion de contenus haineux ou terroristes. Suite à l'appel de Christchurch, Amazon, Google, Microsoft, Twitter et ainsi que le moteur de recherche français Qwant ont diffusé un communiqué ans lequel ils s'engagent à des actions à la fois collaboratives et individuelles. Outre la limitation des « livestreams », ils prônent un partage de données relatives aux contenus délictueux et le développement de technologies facilitant leur détection.

**Les biais de la détection**

Cependant, le manque de transparence et le *modus operandi* de ces algorithmes sont un point sensible. Les réactions, observées face aux contenus à forte charge émotionnelle, et le fonctionnement des algorithmes visant à produire le plus de flux possible, expliquent notamment que des sujets ou vidéos violentes puissent rapidement devenir des sujets « tendance ». Il suffit qu'elles aient été vues ou suffisamment partagées dans un temps assez court pour que les algorithmes prennent le relais des internautes et assurent une diffusion plus large encore.

Autre point délicat, les failles dans les systèmes de modération. Facebook assure que les contenus terroristes sont détectés à 99 % par ses services et que 50 % d'entre eux sont supprimés sous deux minutes.

**Mais les algorithmes détectent-ils tout ?**

Même avec le secours d'une intelligence artificielle éduquée, le problème reste épineux. Cette éducation se fait sur la base de l'analyse de données massives permettant de distinguer ce qui peut être diffusé de ce qui doit être restreint.

À titre d'exemple : alors que les décapitations sont relativement simples à définir, la question se complexifie lorsque l'on aborde le sujet des armes à feu. Pour un système il est difficile de différencier un film de guerre d'une vidéo faisant l'apologie d'une tuerie. En outre, dans le cas de Christchurch, des trolls ainsi que des effets graphiques et des filtres pour modifier l'image ont trompé ce système de contrôle.

L'option alternative est de recourir à un contrôle humain. Mais, dans le cas de YouTube par exemple, il faut environ 70'000 emplois à plein temps pour visionner les quelques 430'000 heures de vidéos posées quotidiennement sur la plate-forme. Un effort financier et humain impensable pour les petites plates-formes.

Malheureusement, ce constat n'a pas échappé à de nombreux sites terroristes qui désormais évitent les grandes plates-formes.

**L'acteur 3 : L'opinion publique et les internautes**

L'opinion publique est le troisième acteur sur cette scène délicate.

Les internautes participaient massivement aux signalements de contenus haineux, violents ou à caractère terroriste, même si dans le cas de Christchurch aucun internaute n'a signalisé le livestream.

Cette incohérence révèle, d'un côté, la volonté de l'opinion publique d'une mise en place rapide de contrôles des contenus circulant sur Internet, alors qu'une autre partie de la population, confrontée à ces contenus, n'agit pas quand elle pourrait le faire. Pourquoi cette inhibition ?

L'absence de signalement lors de la diffusion de la tuerie en Nouvelle-Zélande est cohérente avec la banalisation de la violence : la présence de plus en plus fréquente de contenus violents et radicaux tend à les rendre plus acceptables en inhibant la capacité de réaction.

Néanmoins, le poids des internautes est sous-estimé. S'ils commencent à éviter les plates-formes diffusant des contenus haineux, ce serait l'équilibre économique des entreprises qui serait alors en jeu. Toutes les structures auraient intérêt à intégrer la volonté des internautes dès leur conception, intégrant des procédures de modération efficaces.

Ainsi, si l'appel de Christchurch marque une nécessaire prise de conscience des États et des entreprises, il semble ignorer un certain nombre de problèmes structurels. Sa mise en œuvre est loin d'avoir trouvé une voie entre liberté d'expression et risque de censure, entre business model et protection du public, mais aussi entre raison et appétence malsaine pour le paroxystique.

**Source**
*Siècle digital*
Geneviève Fournier
14 octobre 2019

# 5. La lutte des réseaux sociaux face au terrorisme diffusé en direct

*L'acte criminel de la semaine dernière en Allemagne remet le problème du partage des vidéos sur le devant de la scène.*

Mercredi de la semaine dernière, un allemand de 27 ans tirait sur des passants devant la synagogue de Halle en Allemagne de l'est, ainsi que devant un restaurant turc. Le tireur s'est filmé en direct pendant qu'il commettait ces actions sanglantes. La vidéo durait 35 min, et une partie a pu être diffusée en direct sur Twitch, plateforme de streaming. Comme le rapporte un article dans Vice il semblerait cependant que le partenariat entre les plateformes de partage en partenariat avec les gouvernements, aient porté leurs fruits, évitant ainsi une rediffusion massive de la vidéo, comme ce fut le cas par le passé.

**Une lutte compliquée et sordide**

La diffusion de l'attaque aura permis à 2200 personnes de la visionner. Des copies continuent de circuler sur le réseau Telegram, et 4chan, selon Vice. Néanmoins, le phénomène a été contenu sur des plateformes comme YouTube ou Facebook, ce qui n'a pas toujours été le cas.

En mars, lors de l'attaque de deux mosquées dans la ville de Christchurch, l'auteur des faits avait lui aussi filmé son intervention meurtrière en direct. Malheureusement, la vidéo avait eu le temps d'être visionnée 200 fois en temps réel, avant que quelqu'un ne la signale à la plateforme. Le premier signalement avait eu lieu 12 min après la fin de la vidéo en direct, soit 29 minutes exactement après le début de la tuerie. La vidéo avait ainsi atteint les 4000 vues, et des internautes avaient trouvé le temps de faire des copies pour les éditer sur d'autres sites peu recommandables. À l'image de 8chan, réputé pour accueillir anonymement des commentaires homophobes, racistes, sexistes etc. Le site se compose de beaucoup de personnes déjà exclues de 4chan pour leurs propos extrémistes en tout genre. Google avait d'ailleurs supprimé depuis 2015 le nom de domaine 8chan de ses résultats de recherche, car le site contenait des éléments pédopornographiques.

Ainsi, même si Facebook avait bien entendu supprimé le profil du tueur de sa plateforme ainsi que d'Instagram, le mal était fait. D'après les chiffres rapportés par Vice, en 24h, pas moins de 300 000 copies étaient de nouveau en circulation sur Facebook. Suite à cet évènement choquant et outrageant, avait eu lieu le « Christchurch Call », appel notamment rédigé par la femme Premier ministre néo-zélandaise, Jacinda Ardern et du Président Emmanuel Macron. Cet appel lancé deux mois après le drame en Nouvelle-Zélande, visait à mettre en place des mesures alliant les autorités gouvernementales et les sociétés tech pour éliminer les contenus terroristes et extrémistes sur les plateformes numériques. Le rapport prenait soin de rappeler la nécessité de veiller à préserver la liberté d'expression sur internet, mais précisait que celle-ci ne pouvait encadrer des contenus délibérément violents et/ou terroristes.

C'est ainsi qu'a été créé le GIFCT (Global Internet Forum to Counter Terrorism), organisme indépendant dont l'objectif est de lutter contre l'exploitation des plateformes digitales par le terrorisme et l'extrémisme. Le groupe se constitue de membres représentants de Facebook et ses filiales Instagram et WhatsApp, Google et sa filiale YouTube, Microsoft, Twitter, et 26 États depuis mai 2019. Dans les mois qui ont suivi Amazon, LinkedIn ainsi que d'autres collaborateurs spécialisés en nouvelle technologie ont rejoint l'organisme indépendant, comme le rapportait Facebook le 23 septembre dernier sur sa page Newsroom. Le communiqué rappelait les objectifs établis : analyse en temps réel des contenus, mise en place d'algorithmes ou de signalements permettant de détourner les utilisateurs susceptibles de diffuser ce genre de contenus. Il était également précisé que leur nouvel objectif était de « prévenir » ce genre d'action.

Durant l'attaque de mercredi dernier, ces objectifs ont été mis à l'épreuve. Après avoir supprimé la vidéo du tueur allemand, Twitch a ensuite partagé les données de la vidéo concernée converties sous forme de hash (empreinte unique faite à partir des données fournies, utilisée pour les mots de passe notamment) au GIFTC. Ces données ont ensuite pu être transmises aux autres plateformes

membres, et éviter ainsi la propagation de la vidéo meurtrière.

Celle-ci a cependant été diffusée sur d'autres plateformes, plus petites, qui ne communiquent pas aussi facilement entre elles, ainsi qu'avec les plateformes déjà citées. Adam Hadley, directeur de Tech Against Terrorism, entrée en partenariat avec GIFTC, explique qu'établir une réelle coordination avec ces plus petites structures reste très difficile, mais constitue la clé pour une lutte efficace envers les terroristes et extrémistes qui emploient ces canaux.

D'après Brian Fishman, à la tête de l'équipe « contreterrorisme » de Facebook, c'est un objectif difficile à mener, car certains collègues ne « veulent pas jouer le jeu » déclare-t-il à Vice, et c'est une chose qu'il est nécessaire de réaliser pour avancer précise-t-il.

Autre problème également, qui n'est pas cité dans l'article, la fascination probante de l'être humain pour le sordide, et le voyeurisme. Voyeurisme, qui par ailleurs, fait le succès des réseaux sociaux et des plateformes vidéo. Il semblerait que les criminels, terroristes, extrémistes capables de passer à l'acte et de filmer les violences commises profitent de la perversité d'un système, dont il est désormais difficile de se libérer. Les citoyens peu responsables capables de véhiculer ces crimes, pire, de les cautionner, justifient par la même occasion la création d'organismes tels que le GIFCT, qui sous couvert de vouloir faire de la prévention, pourrait risquer d'entrer dans un système de surveillance accru, faisant fi, malgré lui, de la liberté individuelle déjà bien égratignée depuis plusieurs années par les entreprises numériques.

**Source**
*Le Parisien*
Nicolas Dixmier
17 juillet 2019

## 6. Affaire Bianca Devins : pourquoi les réseaux sociaux n'ont-ils pas empêché la diffusion du meurtre ?

L*es réseaux sociaux rencontrent toujours de nombreuses difficultés à empêcher la propagation des photos du corps de l'adolescente publiées par son meurtrier.*

L'adolescente de 17 ans avait plus d'une centaine de milliers d'abonnés sur Instagram et venait d'être diplômée. Instagram

Et si aucune leçon n'avait été tirée de l'attentat de Christchurch ? Dimanche 14 juillet, en Australie, Bianca Devins, une jeune instagrameuse de 17 ans, a été tuée par Brandon Clark, un homme qu'elle avait rencontré sur le réseau social il y a quelques mois. Il a ensuite publié les photos du corps de la jeune adolescente sur Discord, une plateforme de discussion pour joueurs de jeux vidéo, avant de les mettre sur son propre compte Instagram.

Mais alors que le compte du meurtrier a depuis été supprimé, les photos du meurtre de Bianca Devins continuent de circuler sur les réseaux sociaux comme Twitter, Instagram, Discord ou encore 4chan, un forum habitué à ce genre de polémiques. Même si de nombreux internautes veulent empêcher la diffusion de cette photo en en postant d'autres plus positives de la jeune fille sur le hashtag #RipBianca, elle reste toujours trouvable en suivant certains comptes Instagram qui la partagent.

**Le risque d'une atteinte à la liberté d'expression**

Derrière ce meurtre se cache le problème de la diffusion des contenus violents sur les réseaux sociaux qui deviennent rapidement viraux. Depuis dimanche, Instagram est vivement critiqué pour

ne pas avoir retiré ses images assez rapidement. En mars dernier, Facebook avait été largement critiqué après la diffusion de la tuerie de Christchurch par son auteur, en direct. Mais est-il vraiment possible d'empêcher la diffusion de ces contenus ?

« Aujourd'hui, il est humainement impossible de vérifier tout le contenu mis en ligne à la seconde, explique Valère Ndior, professeur de droit numérique à l'université de Bretagne Occidentale. Pour empêcher la diffusion de contenus violents, les plateformes peuvent bloquer des hashtags, mais portent ainsi atteinte à la liberté d'expression. Par exemple, si c'était le cas dans l'affaire Devins, même un tweet manifestant de l'affection à la famille de la victime serait immédiatement bloqué. »

**Une intelligence artificielle pas encore au point ?**

Alors pourquoi ne pas configurer des algorithmes pour qu'ils soient capables de reconnaître automatiquement tout contenu violent ? « Comment faire pour que le programme différencie un cadavre d'un corps étendu ? Une scène de film d'un réel acte violent ? En cas de doutes, l'algorithme va retirer tout contenu qu'il ne reconnaît pas et cette suppression automatique n'est ni efficace ni souhaitable » estime Valère Ndior.

« L'intelligence artificielle capable de reconnaître automatiquement des scènes de violences existe, mais elle n'est pas encore au niveau de l'intelligence humaine » résume Yves Poullet, professeur émérite de droit numérique à l'université de Namur. « Par exemple, une photo ne peut être acceptable qu'en fonction de son commentaire » qui ne peut être compris que par quelqu'un, explique-t-il pour illustrer le besoin d'une intervention humaine dans la régulation de ces contenus violents.

De plus, même si l'algorithme peut apprendre de lui-même grâce à du data mining, l'analyse d'un grand nombre de données, il suffit qu'un contenu choquant soit modifié pour que le programme ne l'identifie plus en tant que tel et le laisse donc passer. Et même si ce dernier était en mesure de dépasser ces modifications, le contenu circulerait toujours dans les conversations privées des internautes, zones auxquelles l'algorithme n'a pas accès sous peine d'atteindre la liberté d'expression.

**Une question de responsabilités**

Puisqu'il est donc impossible d'empêcher la diffusion de ces contenus violents sans entraver la liberté d'expression, les États ont décidé de légiférer pour responsabiliser directement les plateformes qui diffusent ces contenus. C'est le cas de la France avec la loi Avia dite contre la haine en ligne, mais aussi précédemment de l'Allemagne et de l'Australie.

En essence, ces lois disposent que si les plateformes ne retirent pas des contenus signalés par les autorités et des utilisateurs avant un certain laps de temps, celles-ci peuvent être sanctionnées d'une amende sur leur chiffre de l'affaire annuel qui peut aller jusqu'à 10 % en Australie. Mais « ces sanctions sont avant tout symboliques compte tenu des milliards qu'engendrent ces entreprises » relativise Valère Ndior.

Cependant, en faisant cela, l'autorité publique ne fait que se déresponsabiliser en partie de la régulation de ces contenus. « On remet à des opérateurs privés d'appliquer la loi et d'être des censeurs ou pas, de décider ou non de supprimer le contenu signalé » critique Frédéric Laurie, avocat et maître de conférences à l'Université d'Aix-Marseille.

La diffusion des contenus violents ne dépend donc pas que des plateformes, mais repose « sur un équilibre précaire entre la responsabilité des utilisateurs, de la plateforme et des autorités » comme le résume Valère Ndior. Un problème plus humain que technique, finalement.

**Source**
*Le Figaro*
Harold Grand
20 décembre 2018

## 7. Pédophilie, antisémitisme : les gros problèmes de modération de TikTok, l'application préférée des ados

*L'application chinoise, qui permet de se filmer en train de danser et chanter, est envahie de contenus conspirationnistes et pédopornographiques.*

Si vous vous connectez sur TikTok, vous pourrez trouver des vidéos de karaoké humoristiques, des sketchs, mais aussi des propos néo-nazis. L'application chinoise, qui a récemment racheté son concurrent Musical.ly, est extrêmement populaire auprès des adolescents. Elle revendique aujourd'hui plus de 600 millions d'utilisateurs actifs par mois. Problème: son développement soudain s'accompagne de la prolifération de contenus violents, que son propriétaire, le chinois ByteDance,

peine à endiguer. Une enquête du média américain Vice, publiée mardi, démontre par exemple le nombre important de comptes liés à l'idéologie néonazie. Certains affichent clairement leur soutien à l'Attomwaffen, un groupuscule américain responsable de plusieurs meurtres de personnes juives aux États-Unis. Sur TikTok, on peut aussi taper librement des mots-clés du type #SiegHeil (un slogan nazi), ou consulter des vidéos de jeunes en train de faire le salut nazi.

Le problème est d'autant plus grave que la majorité des utilisateurs de TikTok sont très jeunes. « C'est tout simplement choquant et dangereux, surtout quand on voit le nombre de jeunes qui se sont radicalisés en ligne et ont basculé vers la violence ces dernières années », souligne Heidi Beirich, chargée de projet à la Southern Poverty Law, une association américaine de lutte contre la haine et l'exclusion, interrogée par Vice.

**« Nouvel espace de propagande »**

La France, où 38 % des 11-14 ans ont un compte sur TikTok selon les chiffres de l'association Génération numérique, n'est pas épargnée. « TikTok s'est étendu très rapidement au niveau mondial, très peu de parents savent que leurs enfants sont exposés à ce type de contenus », estime Tristan Mendès-France, enseignant au Celsa spécialisé dans les nouveaux usages numériques. Plusieurs personnalités de l'extrême droite américaine, proche du conspirationniste Alex Jones, ont par ailleurs migré vers TikTok et ont trouvé en cette plateforme un « nouvel espace de propagande idéal, peu utilisé par les structures politiques traditionnelles », précise le chercheur.

Les problèmes de TikTok ne se limitent pas aux contenus liés à l'extrême droite et aux discours de haine. En novembre dernier, la police nationale avait alerté sur les risques de pédophilie sur l'application. Via un tweet, elle appelait les parents à porter plainte si leurs enfants étaient la cible de « propositions sexuelles mal intentionnées ». En début d'année, des médias chinois ont aussi mis en garde les jeunes utilisateurs sur l'usage qu'ils faisaient de la plateforme. Le *South China Morning Post* affirmait par exemple avoir repéré des individus faisant des avances sexuelles sous des vidéos de jeunes en train de danser. En juillet, le gouvernement indonésien a même décidé de bloquer temporairement l'application pour « pornographie et contenu inapproprié ». Le youtubeur français Roi des Rats, spécialisé dans la dénonciation d'abus sur les réseaux sociaux, a de son côté détaillé dans une vidéo publiée début novembre, et vue plus de 2,5 millions de fois, les dérives de la plateforme. Il y dénonçait notamment l'existence de comptes d'échanges d'images pédoporno-graphiques.

« La sécurité des utilisateurs est un défi auquel notre industrie est confrontée à l'échelle mondiale, mais nous prenons cette question très au sérieux et réagissons rapidement », a commenté TikTok auprès du *Figaro*. « Nous continuerons d'améliorer nos efforts de modération et de mettre en place d'autres mesures de protection afin de réduire au minimum les risques d'utilisation abusive. »

À mesure que le nombre de ses utilisateurs explose - l'application était en tête des services les plus téléchargés sur smartphone au premier trimestre 2018 selon le cabinet spécialisé Sensor Tower - Tik Tok fait face à un vrai problème de modération de ses contenus, sans véritable réaction adéquate. Elle a tout de même mis à jour ses paramètres de confidentialité en juin. Il est désormais possible de partager des vidéos « en privé », seulement avec des abonnés approuvés par l'utilisateur, et de supprimer son compte à tout moment. Récemment, TikTok a aussi fait savoir qu'elle allait augmenter ses effectifs de modérateurs humains pour les faire passer de 6000 à 10'000 personnes. « Ces informations restent très floues. On ne sait pas vraiment comment cette modération fonctionne ou si les modérateurs parlent anglais ou français », commente Tristan Mendès-France. « Ce qui est sûr, c'est qu'il faut être extrêmement prudent. Même si l'application est officiellement interdite aux moins de 13 ans, beaucoup d'internautes plus jeunes y sont actifs. » Le chercheur préconise aux parents dont les enfants sont présents sur la plateforme de vérifier si leur compte est privé et si l'accès aux commentaires a été limité dans les paramètres de confidentialité.

**Source**
*SciencePost.fr*
Yohan Demeure
2 novembre 2019

## 8. Cette IA scrute le Twitter chinois pour sauver des centaines de vies du suicide !

*Une intelligence artificielle (IA) créée aux Pays-Bas surveille le réseau social chinois Weibo. L'objectif ? Repérer les messages de suicide et sauver les victimes. Plusieurs centaines de vies ont ainsi déjà été sauvées grâce à ce système.*

**Plus de 700 rescapés en 18 mois !**

Dans un article publié le 9 novembre 2019, la BBC a donné l'exemple de Li Fan, 21 ans, étudiant à l'Université de Nankin (Chine). Après la Saint-Valentin, l'intéressé a posté un message sur Weibo, l'équivalent chinois de Twitter. Avant de perdre conscience, Li Fan disait ne plus vouloir continuer et avoir l'intention d'abandonner. Le jeune homme était endetté, venait de se disputer avec sa mère et souffrait de dépression sévère.

A 8000 km de Nankin, le message de Li Fan est détecté par un ordinateur installé à Amsterdam (Pays-Bas). Un système basé sur l'intelligence artificielle a signalé le post et déclenché une vague de réactions en Chine. Finalement, la police locale a été prévenue et l'étudiant a pu être sauvé de la mort.

Ce genre d'histoire peut paraître incroyable mais, pour les personnes à l'origine du système, il s'agit seulement d'un succès parmi tant d'autres ! En effet, les chercheurs du Tree Hollow Rescue Movement (THRM) de l'Université libre d'Amsterdam ont déjà permis de sauver plus de 700 vies en 18 mois ! Leur système a été spécialement conçu pour détecter les messages de suicide publiés sur Weibo.

**Une portée limitée**

Ce système boosté à l'IA a été pensé pour donner une note allant de 1 à 10 aux messages, en fonction du risque de suicide. Par ailleurs, pas moins de 600 volontaires participent à ce projet en Chine. Lorsqu'un message sur Weibo reçoit une note de 6 ou plus, les volontaires sur place sont prévenus et tentent de contacter la famille ou la police. Néanmoins, selon Huang Zhisheng, chercheur en IA à l'origine du projet, le système ne peut sauver qu'une ou deux vies par jour en Chine, soit environ une dizaine par semaine.

L'expert a indiqué que Weibo limitait l'utilisation des robots d'exploration sur le Web. Ainsi, les chercheurs peuvent collecter seulement 3 000 entrées par jour sur le site. C'est donc pour cette raison que le système se focalise sur les cas les plus urgents, au grand regret de son créateur. De plus, celui-ci peut seulement agir dans l'urgence. En effet, chaque personne sauvée a besoin d'un suivi que l'IA ne sera pas capable d'assurer.

Parfois, les réseaux sociaux eux-mêmes développent leur propre dispositif anti-suicide. C'est le cas de Facebook qui expliquait fin 2017 vouloir démocratiser son système d'IA afin de détecter plus rapidement les messages postés par des utilisateurs suicidaires.

**Source**
*lepoint.fr*
Louis Chahuneau
27 février 2019

# 9. Les modérateurs de Facebook finissent traumatisés et complotistes

C'est l'une des nombreuses entreprises à qui Facebook a délégué sa modération. Son nom ? Cognizant. Elle emploie plusieurs milliers de contractuels pour contrôler chaque jour les contenus du réseau social aux 2,3 milliards d'utilisateurs. Dans une vaste enquête menée auprès d'une douzaine de collaborateurs, le site spécialisé The Verge détaille à quel point les recrues finissent traumatisées par l'expérience. Contrat précaire, isolement, travail surveillé, exposition continue à du contenu choquant laissent les salariés dans un piteux état. « On les incite à ne pas discuter du poids émotionnel que leur travail leur impose, même avec leurs proches », explique notamment le journaliste.

Les 1 000 contractuels employés dans le centre de Phoenix (Arizona) disposent d'une formation de quatre semaines au bout de laquelle ils ne sont pas sûrs d'être embauchés. Leur salaire ? 28 800 dollars par an, quand les salariés de la maison mère de Facebook tournent autour de 240 000 dollars annuellement. Dans son enquête, The Verge raconte la pression mise sur l'ensemble des modérateurs qui contrôlent chaque jour 8 000 messages, photos, vidéos diffusés sur le réseau social. Résultat, un salarié explique passer environ 30 secondes par post pour atteindre 400 traitements par jour. The Verge raconte comment la politique de Facebook en termes de liberté d'expression s'avère parfois contradictoire, ce qui place les modérateurs dans des situations aberrantes. Ainsi, un message comme « Les personnes autistes devraient être stérilisées » ne sera pas retiré, alors que « Je déteste tous les hommes » viole la politique du site. Et attention aux erreurs d'appréciation : les modérateurs qui prennent trop de « mauvaises décisions » risquent le licenciement.

**Sexe, drogue et conspirationnisme**

En plus de la difficulté du travail, les salariés sont entravés pendant leur temps libre : une heure

de pause divisée en trois tranches et une pause de neuf minutes consacrée au « bien-être »... qu'il ne faut pas utiliser pour aller aux toilettes ni pour prier. Cocasse quand on apprend que des centaines d'employés hommes se partagent un urinoir et deux toilettes.

Face à cette pression constante et aux atrocités publiées sur Facebook, certains salariés finissent à bout de nerfs. Ainsi, The Verge révèle comment des modérateurs de Cognizant ont été retrouvés en pleine relation sexuelle dans les cages d'escalier, le garage, et même la salle d'allaitement réservée aux mères. D'autres racontent fumer régulièrement du cannabis pour décompresser, y compris pendant le travail. Enfin, la plupart s'adonnent à la pratique de l'humour noir pour extérioriser leur mal-être, quitte à enchaîner les blagues racistes ou sexistes ou les allusions au suicide.

Devant la prolifération des théories conspirationnistes en ligne, plusieurs contractuels ont avoué douter de plus en plus des versions officielles. Un salarié se balade dans le bâtiment en expliquant que la Terre est plate. Un ancien raconte qu'il a commencé à se poser des questions sur certains aspects de l'Holocauste. Un autre, qui dort avec une arme à feu, a déclaré que l'attentat du 11 septembre 2001 n'était sûrement pas une attaque terroriste.

L'anxiété qui règne dans l'entreprise ne laisse pas les anciens modérateurs indemnes. Plusieurs racontent souffrir de syndromes de stress post-traumatique et de crises d'angoisse en public : « Je ne pense pas qu'il soit possible de faire ce travail et de ne pas en sortir avec un trouble de stress aigu », explique un employé.

**Face au scandale, Facebook réagit**

À la suite de la publication de l'enquête, Facebook a réagi via son vice-président des opérations mondiales, Justin Osofsky. La firme américaine a expliqué qu'elle allait désormais mettre en place des « contrats clairs » avec ses sous-traitants, qui exigeront des suivis psychologiques, des examens réguliers sur leurs performances et des visites régulières dans leurs locaux.

**Source**
afp
13 mai 2020

## 9.1. Facebook va verser 52 millions de dollars pour pallier aux traumatismes de ses modérateurs

*Une plainte avait été déposée devant un tribunal californien par une ancienne modératrice, qui a développé un syndrome de stress post-traumatique. Selon l'accord signé avec le groupe américain, les modérateurs vont recevoir au moins 1000 dollars.*

Facebook a accepté de payer 52 millions de dollars (47,9 millions d'euros) aux modérateurs de contenus en guise de compensation pour les problèmes de santé mentale que leurs tâches peuvent provoquer, ont annoncé mardi deux cabinets d'avocats ayant conseillé les plaignants dans le cadre d'une action de groupe en justice.

Ils reprochent au réseau social planétaire de ne pas protéger correctement ses employés (directs ou via des sous-traitants) chargés de retirer les contenus qui enfreignent les règles de la plateforme.

**Visionnage de contenus violents**

La plainte originelle avait été déposée devant un tribunal californien en septembre 2018, au nom de Selena Scola, une ancienne modératrice qui affirmait avoir développé un syndrome de stress post-traumatique après 9 mois passés à regarder régulièrement des images violentes.

« Tous les jours, les utilisateurs de Facebook diffusent des millions d'images ou de vidéos en direct d'abus sexuels sur des enfants, de viols, de torture, de bestialité, de décapitations, de suicides et de meurtres », relatait la plainte.

« Pour maintenir une plateforme aseptisée, maximiser ses profits déjà conséquents et soigner son image publique, Facebook se repose sur des personnes comme Selena Scola - les «modérateurs de contenus » - pour visionner ces posts et retirer tous ceux contraires à ses règles. »

**Soutien psychologique et enveloppe d'au moins 1000 dollars**

Selon l'accord signé avec le groupe américain, plus de 11 000 modérateurs de Facebook aux États-Unis, anciens et actuels, vont recevoir au moins 1000 dollars (près de 921 euros) chacun. Ceux qui ont été diagnostiqués avec des troubles spécifiques percevront des sommes supplémentaires pour payer leurs frais médicaux (jusqu'à 50 000 dollars).

« Nous sommes reconnaissants aux personnes qui font ce travail important pour faire de Facebook un environnement sûr pour tout le monde », a réagi Facebook, sans reconnaître les allégations de la plainte. « Nous nous engageons à leur fournir les soutiens supplémentaires prévus

par cet accord et plus à l'avenir. »

L'accord prévoit en effet que Facebook et ses sous-traitants fournissent aux modérateurs des sessions de soutien psychologique avec des thérapeutes assermentés et de meilleurs outils pour améliorer leurs conditions de travail.

**Les conditions de travail de sous-traitants déjà égratignées**

« Nous sommes très contents que Facebook ait travaillé avec nous pour créer ce programme sans précédent pour aider ceux qui accomplissent des tâches inimaginables il y a encore quelques années », a déclaré Steve Williams, du cabinet Joseph Saveri.

En 2019, des enquêtes publiées par le site spécialisé The Verge avaient alerté sur les conditions de travail désastreuses de modérateurs employés par Cognizant, qui a depuis arrêté cette activité. Dans la foulée, Facebook avait exigé que ses sous-traitants paient mieux leurs employés et leur fournissent un accompagnement psychologique.

**Source**
*The Conversation*
Thibault Grison
9 septembre 2021

# 10. IA et modération des réseaux sociaux : un cas d'école de « discrimination algorithmique »

Ces dernières années, les entreprises détentrices des réseaux sociaux numériques (comme Google, Facebook, Twitter ou ByteDance, l'entreprise détentrice du réseau social TikTok) ont beaucoup investi dans le recours à l'IA et à l'apprentissage automatique pour décupler leurs capacités de repérage et de modération des contenus illicites sur le web. La montée de la haine en ligne ou encore l'emballement médiatique autour des fake news ont instauré un climat médiatique parfois hostile aux contenus diffusés sur les réseaux sociaux et une panique morale pour réguler le web à tout prix. Entre bras de fer législatifs et marronniers médiatiques, la manière de modérer les réseaux sociaux numériques est passée d'un questionnement technique opéré à huis clos par les entreprises à un enjeu éminemment politico-publique.

Toutefois en parallèle de ce recours plus massif à l'IA, certaines communautés en ligne disent avoir fait l'objet d'une vague de censure abusive de la part des plates-formes. Comment le recours à l'IA dans la modération des réseaux sociaux engendre-t-il des discriminations à l'encontre de certaines catégories protégées ?

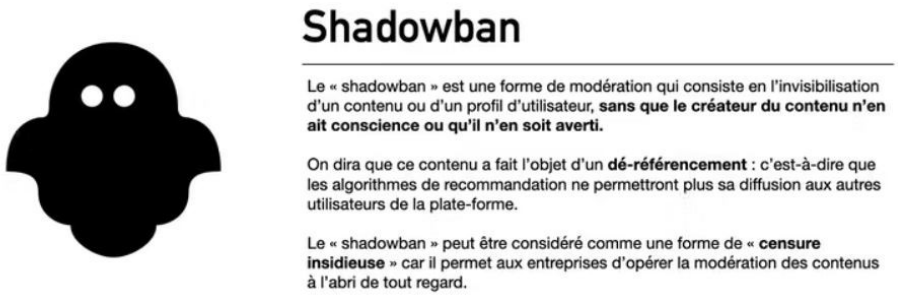
**Lutter contre la prolifération des contenus illicites en ligne, mais à quel prix ?**

Aujourd'hui, nul ne sait vraiment comment la haine en ligne est concrètement modérée sur les réseaux sociaux numériques.

On sait que des entreprises comme Facebook et Twitter emploient des modérateurs dans des sociétés de sous-traitance à l'étranger dans des conditions de travail parfois douteuses, mais aussi qu'elles possèdent leurs propres équipes de modérateurs pour gérer les contenus illicites et le référencement des publications en interne, et enfin qu'une partie de cette modération est de plus en plus déléguée à des algorithmes de *machine learning*, chargés de « nettoyer » les plates-formes de tous les contenus qu'elles auront jugé indésirables conformément aux règles fixées par les chartes et conditions d'utilisation des plates-formes.

Le travail de modération sur les réseaux sociaux numériques est donc à la fois réalisé par des humains et de manière automatisée, sans que l'on connaisse avec certitude les modalités d'articulation de ces méthodes, ainsi que la part de contenus traités exclusivement de manière automatique.

Or la délégation du travail de modération à l'IA semble aller de pair avec la discrimination à l'encontre des minorités sexuelles, de genre et de race. En effet ces dernières années, de nombreux utilisateurs ont tenté d'alerter les entreprises au sujet des injustices dont ils étaient victimes sur leurs plates-formes respectives. Ces injustices pouvaient prendre les formes suivantes : suppressions de comptes, censure automatique de posts, démonétisation de vidéos ou encore dé-référencement de contenus, via le phénomène de *shadowban*.




Le *shadowban*, une des formes de modération de contenus en ligne. Thibault Grison

Les formes de censure sont donc multiples et divergent d'un réseau social numérique à l'autre.

**De nombreux exemples de discriminations liées à la modération en ligne**

Prenons l'exemple des utilisateurs et utilisatrices ouvertement LGBT+ pour illustrer notre point. En 2019, le mouvement « SEO Lesbienne » avait tenté de créer un compte Facebook comportant le mot « lesbienne », en vain. Ce nom d'utilisateur était automatiquement refusé par la plate-forme pour le motif suivant : « Il comporte des mots qui ne sont pas autorisés sur Facebook ».

Plus récemment sur Facebook et Twitter, des comptes de militants LGBT ont été temporairement suspendus en raison de la mention de leur orientation sexuelle dans leur bio (Twitter) ou descriptions de leur photo de profil (Facebook) : l'IA assimilait la mention de mots-clés comme « pédé » ou « gouine » a du contenu nécessairement haineux sans prendre en compte le contexte d'énonciation et la réappropriation militante de ces termes.

Sur TikTok, une étude menée par le think tank australien ASPI (Australian Strategic Policy Institute) a montré que selon la langue utilisée par l'utilisateur des réseaux sociaux numériques, certains hashtags LGBT+ étaient déréférencés de la plate-forme. Citons-en quelques exemples : يسنج #لوحتملا soit « transgenre » en arabe, #ягей pour « je suis gay » en russe, #Intersex pour « intersexe » en anglais ou encore #gej pour « gay » en bosnien et estonien. Ces quelques exemples constituent une des facettes de ce à quoi peut ressembler la discrimination algorithmique dans la modération des réseaux sociaux numériques.

On appellera discrimination algorithmique toute discrimination causée par un recours aux algorithmes. Il s'agit, en d'autres termes, de l'expérience de la discrimination vécue par les utilisateurs de ces services automatisés. Mais quelle en est la cause ?

**À l'origine, les biais algorithmiques**

La discrimination algorithmique résulte de l'introduction de « biais » au moment de la conception des algorithmes. Ces biais consistent en la transposition d'observations générales (souvent stéréotypées) ou statistiques en conditions algorithmiques systématiques. Il en existe plusieurs types et ils peuvent apparaître à différents moments du cycle de vie d'un algorithme. Je propose de les réunir en trois grandes catégories.

La première relève de la qualité des données d'apprentissage des algorithmes de *machine learning*. Dans leur écrasante majorité, ces données contiennent déjà des biais discriminants, que les algorithmes vont ensuite reproduire de façon « mécanique ». Par exemple, les algorithmes entraînés à partir de corpus moissonnés sur le web peuvent finir par associer le mot « lesbienne » à des contenus pornographiques et donc à considérer des contenus militants comme problématiques.

La seconde catégorie de biais porte sur ce qu'on appelle tantôt « biais de société », « biais cognitifs » ou encore « biais de stéréotypes » selon les langues et champs de recherche en IA. Il s'agit des représentations biaisées et des impensés des concepteurs (humains) transférés aux machines au moment de la conception.

Enfin, les objectifs de rentabilité et des critères garantissant l'efficacité d'un algorithme peuvent aussi engendrer des biais algorithmiques. Selon l'importance qu'on donnera à un critère (souvent financier) plus qu'à un autre, les algorithmes, pour répondre à cet objectif qui constitue leur raison d'être, pourront engendrer des discriminations collatérales. En guise d'exemple, imaginons un algorithme de recommandation dont la mission principale serait de promouvoir les contenus qui génèrent le plus de trafic en ligne (et donc de revenus pour les entreprises comme Google).

Le mot « lesbienne » renvoyant essentiellement à du contenu pornographique destiné à un public hétérosexuel et masculin ; les algorithmes de recommandation mettront en avant ce type de contenus (parce que reconnu comme populaire) dans les résultats de recherche. La pertinence des résultats de recherche ne coïncide donc pas avec les enjeux de représentation et de visibilisation des différentes orientations sexuelles et identités de genre en ligne, qui n'ont tout simplement pas été pris en compte par les concepteurs. On appellera cette négligence un « biais de variable omise ».

Trop souvent perçus comme neutres et objectifs, les algorithmes (comme toute technologie) peuvent, selon les usages et les concepteurs qui les mettent en service, reproduire des injustices déjà existantes dans la société.

La question qui se pose à présent est celle de la responsabilité. Si des titres de presse comme « L'algorithme anti-haine de Google est raciste envers les noirs » ou « TikTok's algorithm is promoting homophobia » semblent caricaturaux aux yeux des défenseurs de l'IA qui jugent absurde le fait d'imputer la responsabilité de la discrimination aux algorithmes et à leurs concepteurs, le problème de la discrimination demeure.

Pourtant, dès lors que les cas de censure en ligne se multiplient et que ces outils sont toujours déployés, il appartient aux communautés scientifiques et militantes de mettre au jour les effets de ces outils sur les populations et la responsabilité des entreprises à maintenir leur mise en service alors même qu'ils contribuent à la (re)production d'injustices.

De plus, le désir de légiférer contre la haine en ligne ou les fake news ne saurait se faire sans un contrôle des outils automatiques de modération, au risque d'accroître les cas de discriminations algorithmiques. En effet, le projet de loi visant à lutter contre les contenus haineux en ligne portée par la députée LREM Laëtitia Avia a fait craindre un recours plus accru aux algorithmes en raison du nouveau délai de 24 heures imposé aux plates-formes pour supprimer les contenus qu'elles auraient jugé illicites. Plutôt que de proposer une loi pour lutter contre la haine en ligne qui risquerait d'avoir de lourdes conséquences sur la liberté d'expression des personnes que la loi vise à protéger, il conviendrait d'abord de réduire l'impact discriminant du recours aux algorithmes dans les réseaux sociaux en pénalisant les détenteurs des plates-formes lorsqu'ils discriminent des populations.

Les algorithmes interviennent de plus en plus dans nos vies quotidiennes et affectent la manière dont nous recevons l'information et percevons le monde – pensons à réguler la manière dont ils nous régulent, avant de leur donner plus de pouvoir.