Le traitement automatisé des injures

Grégoire BEN-AÏSSA, Thomas BERNS, Tyler REIGELUTH

Résumé :

Cet article propose une description et une analyse des activités automatisées de régulation, sélection et modération des contenus sur les plateformes, en partant pour l'essentiel de l'exemple de Facebook et du cas des discours de haine. De telles plateformes doivent-elles être désormais considérées comme des « institutions discursives » ? Et développent-elles les moyens pour réaliser cet objectif ? En mobilisant la reprise de la théorie austinienne du performatif par Jacques Derrida et Judith Butler, et en confrontant les différents types de normativités – technique, juridique et éthique - à l'œuvre dans l'espace discursif contemporain, nous montrons que c'est l'épreuve et le témoignage du caractère fragile et incertain de toute citation qui pourrait s'estomper.

Abstract:

This article describes and analyzes the automated regulation, selection and moderation of content on platforms, using the example of Facebook and the case of hate speech. Should such platforms now be considered as "discursive institutions"? And do they develop the means to accomplish this objective? By mobilizing Jacques Derrida and Judith Butler's interpretation of the Austinian theory of the performative, and by confronting the different types of normativities - technical, legal and ethical - at work in the contemporary discursive space, we show that it is the experience and testimony of the fragile and uncertain character of any citation that could fade away.

Mots clés :

Algorithmes, Austin, Butler, Derrida, Discours de haine, Facebook, Injure, Insulte, Intelligence artificielle, Itérabilité, Normativité, Performativité, Plateformes, Régulation, Réseaux sociaux.

Key words :

Algorithms, Artificial intelligence, Austin, Butler, Derrida, Facebook, Hate speech, Insult, Iterability, Normativity, Performativity, Platforms, Regulation, Social networks

Figure issue d'une capture d'écran datant du 27 mai 2021 et extraite d'un groupe de discussion sur le cyclisme. Le propos, automatiquement perçu comme homophobe sur la base du traitement algorithmique des discours de haine, a été bloqué pour non-respect des standards de la communauté. Est ainsi exacerbée une tendance plus générale que Judith Butler critique dans *Le pouvoir des mots* conduisant à considérer que le pouvoir injurieux des mots leur serait intrinsèque.

Parallèlement à la surenchère informationnelle sur les plateformes numériques, les stratégies et techniques de recommandation se sont généralisées à tous les niveaux d'interaction. L'extension de la recommandation, comme normativité douce et continue, qui cherche à orienter les comportements des utilisateurs de manière quasi insensible et non coercitive, s'alignait évidemment sur un certain libéralisme (voire libertarianisme) porté par les plateformes de la Silicone Valley. Ce

processus de recommandation s'apparente à un jeu d'interactions entre utilisateurs et systèmes algorithmiques qui s'informent dans des boucles récurrentes. Si le mythe d'une horizontalité et d'une autorégulation des échanges sur les plateformes a pu un certain temps dissimuler la fonction des médiations techniques au sein de ces échanges, le fait que ces plateformes agissent comme de véritables régulateurs de discours ne fait plus de secret. En effet, ces dernières sont appelées non seulement à visibiliser des contenus, à travers un processus de recommandation à géométrie variable, mais aussi à interdire certains contenus, c'est-à-dire à les exclure du jeu de la recommandation. Ce qui devait être un marché auto-régulé du discours où l'indicible se retrouverait spontanément hors-jeu est ainsi devenu une institution discursive, où le partage entre le dicible et l'indicible doit être arbitré *comme* par un tiers. Ce qui est remarquable est non seulement la résurgence de ce tiers qui, comme on le verra, peine à se hisser au niveau de l'institution, mais également le fait que celle-ci converge avec un autre phénomène de fond, à savoir l'automatisation (apparente) des activités humaines. De fait, que certains propos racistes ou sexistes soient exclus ou que certains discours qui enfreignent les règles du processus électoral soient contrecarrés n'est pas étonnant. Ce qui interpelle c'est la mobilisation par les plateformes de systèmes d'intelligence artificielle, développés par leur soin, pour effectuer cette régulation à l'échelle des millions de discours qu'elles hébergent chaque jour. La modération de contenus se pratique à l'échelle et à la vitesse de la contagion discursive : elle ne correspond plus à un moment ou un geste discursif supplémentaire, se déroulant après coup, comme avec la régulation par le droit, mais s'inscrit de plus en plus dans la temporalité même du discours qu'il s'agit de réguler.

Ce texte a vocation à analyser les spécificités d'une telle automatisation de la régulation pour ensuite les questionner sur la base d'une lecture derridienne et butlerienne du performatif. Il ne s'agira donc pas de soutenir que la régulation des discours par les processus automatiques est impossible ou indésirable, mais plutôt de montrer qu'une telle automatisation ne peut dépasser l'échec qui guette, nécessairement, toute régulation alors que les processus automatisés ont sans doute pour principale "vertu" de voiler, voire d'invisibiliser, cet échec. Cet échec est incompressible, ne peut pas être résorbé par le processus de régulation ; il n'est que le déplacement que la régulation produit elle-même et il est lui-même porteur d'effets et d'échos. Il ne s'agit donc en rien de simplement prôner un laisser dire, et donc un laisser faire, mais de rendre visible l'impossible résorption de l'échec d'une régulation par son automatisation. C'est avec une attention particulière aux différentes modalités techniques de ces systèmes algorithmiques que nous tenterons de cerner comment l'échec se joue dans la régulation des discours de haine. Pour se faire, nous présenterons dans un premier temps quelques distinctions techniques pour introduire des cas concrets où celles-ci se mêlent à des modes de justification de la part des plateformes. Nous nous baserons pour l'essentiel

sur l'exemple de Facebook non pas parce qu'il serait le réseau social contemporain par excellence mais au nom de sa relative ancienneté qui en fait un outil bien expérimenté et documenté exprimant ainsi quelque chose de symptomatique sur l'évolution de la prise en charge de la régulation par ces plateformes (Gillepsie 2018). De surcroît, à l'inverse d'autres plateformes plus récentes dont la structuration semble elle-même susciter les propos haineux (ce qui aurait eu pour conséquence de déplacer notre analyse vers cette incitation, là où nous entendons nous limiter à celle de leur contrôle), les échanges sur Facebook nous sont apparus comme plus "naturellement" modérés, en somme plus proches du langage ordinaire. L'enjeu plus explicitement politique auquel cela nous amène est de rendre manifeste le monopole de ces plateformes privées sur certains champs discursifs et leur pouvoir de censure insensible, qui tend alors à s'exprimer comme une simple organisation du visible, bien plus que comme une interdiction.

1. L'émergence des discours de haine sur internet et la justification de leur traitement par algorithme

Les discours de haine sur internet

La littérature scientifique sur le traitement algorithmique des discours de haine porte généralement sur des prototypes d'algorithmes. Elle fournit quelques éléments de définition du discours de haine, dont trois apparaissent centraux : la violence, le ciblage et la nuisance. Un discours de haine est un discours violent, que cette violence soit inhérente à l'énoncé, offensant en lui-même (Putri et al, 2020, p.1) ou en tant qu'il exprime une "opinion" (*idem*) qui rabaisse une certaine catégorie de population, en tant qu'il exprime un sentiment de haine (Martins et al, 2018, p.2) ou encore en tant qu'il incite à la violence (Herwanto, Trisna, 2019, p.1). Un discours de haine est aussi un discours qui cible une certaine personne ou un groupe de personnes (Martins et al, 2018, p.2) sur la base d'une caractéristique spécifique (Putri et al, 2020, p.1; Herwanto, Trisna, 2019, p.1), qu'il s'agisse de la nationalité, de la confession, de l'ethnie, de l'identité de genre ou encore de l'orientation sexuelle. Enfin, la nuisance (blessure, discrimination, menace, etc.) produite par cette violence envers les personnes ciblées en vertu d'une caractéristique spécifique est également un élément courant de la définition du discours de haine. On retrouve par exemple cette idée derrière la notion de "préjudice" (Herwanto, Trisna, 2019, p.1), qui nous situe ainsi clairement dans un cadre conceptuel

qui nous permettra ensuite de convoquer les arguments de Judith Butler dans *Le pouvoir des mots* selon lesquels les discours de haine minent la puissance d'agir des personnes insultées.[1]

Dans ses standards de la communauté, Facebook définit le discours de haine comme une "attaque directe contre des personnes, plutôt que contre des concepts ou des institutions, fondée sur ce que nous appelons les "caractéristiques protégées"...", attaque elle-même définie comme discours violent ou déshumanisant, affirmation d'une infériorité, expression d'un mépris, d'un dégoût, appel à l'exclusion, etc. (voir la page "Discours haineux", *Meta Transparency Center*, n.d. [2022]). Est également mentionnée l'idée d'après laquelle les discours de haine "créent un environnement intimidant et excluant et peuvent, dans certains cas, faire l'apologie de la violence hors ligne", nuisant à la puissance d'agir des personnes ciblées par ces discours et à leur expression non-contrainte ("Discours haineux", *Meta Transparency Center*, n.d. [2022]). En somme, Facebook définit les discours de haine conformément à la littérature que nous avons évoquée : le préjudice est à comprendre dans les termes de l'exclusion ou de l'intimidation, qui viennent le spécifier, la violence renvoie aussi bien aux offenses en ligne qu'à la violence physique qui peut en découler, et les "cibles" sont précisément les personnes appartenant aux populations couvertes par les "caractéristiques protégées".

La mise en place des standards de la communauté comme contractualisation

Pour encadrer les discours de haine, et l'action des algorithmes à leur égard, Facebook a mis en place des "standards de la communauté", règles s'appliquant à tous partout dans le monde et pour tous types de contenus, leur donnant par conséquent, ainsi qu'à l'action des algorithmes, une étendue considérable ("Standards de la communauté Facebook", *Meta Transparency Center*, n.d. [2022]). Ces règles, qui peuvent être considérées comme permettant de contractualiser une certaine police du langage sur la base d'un accord de l'utilisateur requis pour son adhésion à la plateforme, prennent en considération aussi bien les "retours d'utilisateurs" que l'"avis d'experts spécialisés dans des domaines tels que les technologies, la sécurité publique et les droits de l'homme", affirme Facebook. Voulant

[1] Par la suite, nous emploierons de manière relativement indistincte les trois expressions voisines que sont « insulte », « injure » et « discours de haine », et leurs dérivés respectifs. Si ces expressions renvoient à une même réalité, celle de propos qui blessent ou peuvent blesser, elles mettent l'accent sur des aspects différents de cette réalité ; leur triangulation continue donc d'être nécessaire pour couvrir la totalité du phénomène visé, tout en sachant qu'il n'est pas possible – c'est même là un des buts de cet article – de produire une formule qui hiérarchiserait les différents aspects en question. En bref, « insulte » met l'accent, par son étymologie, sur la violence intrinsèque du propos, voire sur la dynamique qu'y inscrirait le locuteur ; « injure » rend compte non seulement de la blessure qui peut en résulter (*injury*) mais aussi bien plus globalement à l'atteinte aux droits d'autrui ; enfin « discours de haine » (et plus précisément « incitation à la haine » à partir d'un discours) met l'accent sur la qualification juridique de ces insultes et injures tout en rendant compte , comme le veut le droit en la matière, de la nécessité d'établir un lien fort entre le propos et son action sur autrui, au point de considérer parfois que l'action est incorporée dans le(s) mot(s).

s'assurer que "tout le monde a voix au chapitre", ces normes "prennent en compte les différents points de vue et croyances, en particulier ceux des personnes et des communautés marginalisées ou négligées". Est proposée une liste non exhaustive des cas dans lesquels les standards de la communauté seraient considérés comme violés, entrainant sanction. Une différence est par exemple établie entre le contenu non autorisé, qui se verra bloqué automatiquement, et le contenu requérant davantage d'informations. Mais la diversité des contenus entrant dans ces cadres, à protéger ou à interdire, depuis la charte en question, est impressionnante : il peut s'agir aussi bien des discours de haine au sens strict (dont la liste est à son tour longue et diversifiée : "Discours haineux", *Meta Transparency Center*, n.d. [2022]) que de la lutte contre les activités criminelles, contre la fraude ou contre le harcèlement, de la gestion de la nudité, de la protection de la vie privée, de la cybersécurité, de la propriété intellectuelle ou encore de la protection des mineurs, etc. Il est par ailleurs à noter que dans ce cadre la version de référence et la version rédigée en anglais (américaine), celle-ci étant la plus à jour et servant de "document principal".

La légitimation de leur action par les plateformes

Facebook souligne que ces standards de la communauté ont été mis en place dans le but de "créer un lieu d'expression qui donne la parole à tous", dans lequel chacun puisse s'exprimer. Les discours de haine portent donc atteinte à un tel lieu d'expression en excluant, en minant les capacités expressives et d'interaction (paisible), des destinataires. Facebook fait également appel à l'idée de dignité en soulignant "Nous attendons de chaque personne qu'elle respecte la dignité d'autrui, et qu'elle ne harcèle pas ni ne rabaisse les autres." Enfin, comme on l'a dit, la mise en place des standards de la communauté supposés guider l'action des algorithmes est mise en avant comme répondant à une demande des utilisateurs et des experts, et ainsi représenter quelque chose comme la contractualisation de volontés individuelles au sein d'une communauté, sinon un sens commun.

Si de telles justifications sont claires, on peut néanmoins se demander pourquoi privilégier le traitement par algorithme des discours de haine plutôt que le recours à la loi, sachant que cette question s'accompagne aussitôt de celle de savoir si la plateforme est légitime à constituer des normes sur le sujet et à en garantir l'application. Dans ce cadre, l'argument phare non seulement dans l'utilisation mais aussi dans le perfectionnement des algorithmes de traitement des discours de haine est bien entendu celui de l'efficacité. Facebook se vante de la détection proactive, c'est-à-dire automatique et préalablement à tout signalement des utilisateurs, de 94.7% des discours de haine ("Discours haineux", *Meta Transparency Center*, n.d. [2022]) - et surtout d'une efficacité aussi discrète et indolore que possible, en apparence du moins (nous devrons y revenir), c'est-à-dire qui entrave le

moins possible le bon développement de l'espace d'expression. Dans le développement de ses algorithmes Facebook se targue d'ailleurs de sa démarche d'"open science". La plateforme semble considérer que cette démarche ait pour mérite de permettre un accès (donc un contrôle) direct des utilisateurs aux algorithmes utilisés, dans le but de désamorcer toute possible critique sur l'opacité concernant l'action et le développement de ces algorithmes. Ce présupposé prend pour acquis, en dépit des quelques pages de vulgarisation et d'explication qu'offre la plateforme à ce sujet, les compétences techniques requises pour effectivement bénéficier de cette publication, pour comprendre et éventuellement critiquer le fonctionnement des algorithmes.

Si le respect des règles de la communauté des utilisateurs sert encore de discours légitimant, dans la régulation des contenus par Facebook la logique de plateforme tend à supplanter ce rapport "communautaire". En effet, le problème pour une plateforme comme Facebook est moins de savoir quel contenu supprimer ou non, mais plutôt de pouvoir faire face à la viralité des discours qui s'y propagent. La plateforme mise alors sur l'automatisation pour assurer l'efficacité de cette régulation, l'idée étant que seul un traitement algorithmique peut suivre le mouvement de cette propagation à une échelle et une vitesse inédite (Gorwa, Binns et Katzenbach, 2020). Idéalement - c'est en tout cas l'horizon promis par ces plateformes aux législateurs et régulateurs publics – ce traitement algorithmique reposera de plus en plus sur des systèmes d'intelligence artificielle capables de prédire la toxicité ou la dangerosité d'un discours avant même qu'il ne produise son effet de nuisance et en parant donc à toute forme de répétition virale. Cela nous invite à regarder, derrière les discours et de légitimation que tient la plateforme sur son action, comment fonctionne la modération algorithmique des discours sur la plateforme à l'heure actuelle.

2. Fonctionnement : traitement algorithmique des discours de haine

Les méthodes

On distingue usuellement deux principales méthodes dans la détection automatique des discours de haine (Vinot, Grabar, Valette, 2003, p.276). D'une part, le filtrage par liste noire qui consiste à filtrer les pages dont l'adresse (URL) fait partie d'une liste préalablement constituée à cette fin, en les bloquant. L'idée est d'empêcher la prolifération des discours de haine en agissant directement à la source. Compte tenu du caractère expansif des ressources disponibles sur internet et de leur fluidité, le principal problème de cette approche est la mise à jour de la liste noire. Il s'agit toutefois d'un mode de régulation qui ne réclame pas nécessairement d'intelligence artificielle et qui

repose sur l'identification de sources jugées problématiques et non de certains discours ou énoncés en tant que tel.

L'autre technique est celle du filtrage par mots qui réglemente l'accès à une page ou publication en fonction de la présence de mots clés. A cette fin, il est davantage utile de se baser sur des "sacs de mots" qui permettent la constitution de syntagmes identificatoires, en intégrant d'autres mots non spécifiques au discours de haine mais y étant souvent associés tels que "envie", "agressions", "désinformation" ou "honte" (Vinot, Grabar, Valette, 2003, p.280). Cette technique utilise généralement d'autres indices linguistiques (caractères, morphèmes ou encore catégories syntaxiques), péri-textuels (sommaire, rubriques ou titres), ou non textuels (nombres ou code HTML), de manière à permettre de distinguer les contenus racistes et les contenus antiracistes portant sur ces derniers (nous devrons revenir sur cette difficulté essentielle de la citation). Cette approche – comme celle de la liste noire d'ailleurs - a le mérite de rendre relativement visible la convention à partir de laquelle un discours sera signalé et de laisser la possibilité à des acteurs, tels que les utilisateurs, des experts sectoriels ou des membres de la société civile de participer à l'identification de ces mots, contenus ou sources. Cependant, cette approche présente également certains angles morts, dans la mesure où elle prend difficilement en compte la polysémie et la diachronie des mots. Elle peut ainsi faire l'impasse sur les phénomènes de créativité linguistique qu'il s'agisse de phénomènes de "réhabilitation des mots" ou de l'usage du verlan, de modifications d'orthographe, d'emprunts à d'autres langues, et autres techniques utilisées, entre autres, dans le but de contourner l'algorithme (Vinot, Grabar, Valette, 2003, p.276).

Les plateformes semblent elles-mêmes conscientes des limites d'une telle entreprise, à l'image de Facebook qui explique " Nous reconnaissons que les utilisateurs partagent parfois des contenus incluant des insultes ou le discours haineux de quelqu'un d'autre pour le condamner ou sensibiliser les autres à son égard. Dans d'autres cas, des discours, y compris des insultes […] peuvent être utilisés de manière autoréférentielle ou de manière valorisante. Nos politiques sont conçues pour permettre ce type de discours, mais nous demandons aux utilisateurs d'indiquer clairement leur intention. Nous nous réservons le droit de supprimer le contenu concerné lorsque l'intention n'est pas claire" ("Discours haineux", *Meta Transparency Center*, n.d [2022]). On entrevoit déjà la manière dont une telle exigence d'explication, de réflexivité ou de métacommunication sur les intentions semble contraire à toute réalité langagière concrète.

Néanmoins, l'approche choisie semble principalement rester celle du filtrage par mots clés, basé sur des éléments et des énoncés "historiquement [utilisés] pour attaquer, intimider ou exclure des groupes spécifiques et souvent [liés] à la violence hors ligne." Facebook précise encore : "en

fonction de nuances locales, nous prenons parfois en considération certains mots ou certaines phrases, comme les termes fréquemment utilisés pour désigner les groupes de CP [caractéristiques protégées].” Cette technique requiert alors une définition toujours plus précise des discours de haine et des injures qu'il s'agit de sanctionner, en les bloquant. Un des défis souvent soulignés dans la littérature est la difficulté à saisir la dimension contextuelle d'un énoncé, précisément parce que le contexte est en grande partie indisponible de par la nature distanciée et techniquement médiée des communications numériques (Gorwa, Binns et Katzenbach, 2020). En fait, l'enjeu n'est pas tant que le contexte “réel” soit indisponible, mais que la communication sur les plateformes modifie le contexte en même temps qu'elle essaie de l'établir. Parallèlement à ces techniques, de plus en plus de systèmes reposent sur des processus d'apprentissage et de prédiction algorithmique pour signaler et éventuellement bloquer ou réduire la visibilité des discours de haine.

Qualification et détection, blocage et sanction

Si, à l'image de la loi, les techniques de la liste noire ou du sac de mots semblent pouvoir faire place à des gestes qu'on peut au moins théoriquement distinguer - la détection (de l'énoncé *potentiellement* injurieux, qu'elle soit automatique ou qu'elle fasse suite à un signalement), la qualification (de cet énoncé *comme* injurieux) et la sanction (blocage ou réduction de la visibilité de l'énoncé qualifié d'injurieux) -, le recours à des algorithmes apprenants rend cette distinction de plus en plus poreuse et dynamique. De plus, là où l'administrateur (à l'image du juge) disposait d'une discrétion, d'une marge d'appréciation et d'interprétation, les nuances propres à l'action de l'algorithme semblent pour leur part se jouer de la sorte : il sera demandé seulement à certains contenus d'être précisés, mais la plupart seront automatiquement bloqués, bien qu'une possibilité de recours subsiste par la suite (“Discours haineux”, *Meta Transparency Center*, n.d. [2022]).

Plus précisément, sera automatiquement supprimé par les algorithmes, et cela en vertu des “politiques” mise en place par Facebook (“Discours haineux”, *Meta Transparency Center*, n.d. [2022] ; “Standards de la communauté Facebook”, *Meta Transparency Center*, n.d. [2022]), tout contenu identifié comme “non-autorisé”, comme celui “qui décrit ou cible négativement des personnes par des injures, où les injures sont définies comme des mots intrinsèquement offensants ou utilisés pour insulter des personnes sur la base des caractéristiques” citées par Facebook (ethnie, nationalité, handicap, religion, caste, orientation sexuelle, sexe, identité de genre...). Mais dans certains cas, les suites de la détection ne sont pas si évidentes et tranchées. Pour certains contenus, davantage d'informations ou de contexte seront réclamés pour que la plateforme puisse s'assurer du respect des standards de la communauté. Il peut s'agir de contenus satiriques, tournant en dérision ou citant dans

une démarche critique les contenus “non-autorisés”. Ce sont ces contenus qui requerront l’explication des intentions de leur auteur, comme nous l’avons évoqué plus haut. Ces contenus permettent d’apercevoir les limites et la mise en difficulté du traitement algorithmique des discours de haine.

L’arsenal des réponses : blocage, suppression, neutralisation

Se pose dès lors la question de la modalité de l’interdiction d’un contenu : s’agit-il de le supprimer *ex post* ou de le bloquer *ex ante* ? Dans le premier cas, un contenu supprimé ne subsiste que par la trace de sa suppression, qui se manifeste par la publication de la justification (ou d’une indication a minima). Le discours agit alors encore comme possibilité, non plus cette fois-ci dans le fait qu’elle ait blessée ou non mais dans la multiplicité d’énoncés blessants que l’on est amené à imaginer face à son absence. Par le fait même d’avoir éteint le jeu de déplacements possibles, un nouveau jeu, plus spéculatif, s’ouvre quant à savoir ce qui a bien pu mériter une telle censure. Par ailleurs, dans de nombreux cas, la trace de la publication originale a été conservée par le publiant sous forme de photo ou de capture d’écran, ce qui en fait un élément de preuve mobilisable dans la dénonciation d’un acte de censure abusif ou insensé, mais aussi ce qui semble souligner le caractère fantasmatique d’une régulation totale des discours, même par algorithme. C’est le cas notamment de publications sur Facebook comprenant le terme “pédale(s)” qui tendent à être retirées de manière abusive car elles constitueraient un discours haineux alors que la mobilisation de ce terme renvoyait effectivement au champ sémantique du cyclisme ou de la guitare, sur des pages (une page dédiée au matériel de guitare) ou associé à des images (une image d’un vélo avec une pédale cassée) qui permettait pourtant de régler sans trop de difficultés l’indétermination du contexte. De manière encore plus problématique, certains militants LGBTQI+ ont vu leurs propos censurés ou leurs comptes temporairement bloqués pour l‘usage d’un terme que la communauté s’est péniblement réapproprié. Aveugle à cette réappropriation, la suppression de ces discours peut s’apparenter, pour ces usages, à une blessure supplémentaire qui renvoie le discours à celui duquel il a justement cherché à s'émanciper (voir la figure à la fin du document).

Dans le cas, le plus emblématique et en apparence le plus efficace, où il s’agit tout simplement d’empêcher la publication d’un contenu, l’interdiction se joue de plus en plus au niveau d’un calcul de la probabilité qu’un discours enfreigne les règles d’usage de la plateforme et constitue un propos haineux. Que l’acte même de la suppression soit automatique ou requiert une validation humaine (une validation humaine qui, étant coûteuse, semble s’effectuer sur le mode le plus “machinal” !),

l'enjeu demeure le même : empêcher la blessure avant qu'elle puisse être constatée. Les stratégies d'évitement dans ces cas consistent généralement à modifier un élément du discours (en écrivant par exemple "p*dale(s)") pour échapper à la détection par le système ayant détecté et supprimé le contenu initial. Notons également que, même dans le cas de ces contenus bloqués, puisqu'un recours est possible, et que le discours peut potentiellement être rendu de nouveau accessible, il doit subsister une trace (non publique donc non accessible) du discours, dans les "logs", fichiers permettant de stocker, au moins temporairement, l'historique des événements advenus sur le serveur, donc d'archiver un certain temps le discours supprimé. Signalons d'ailleurs que la nature et la temporalité d'une telle archive, virtuelle et secrète voire invisible, reste entière et encore faiblement problématisée à l'heure actuelle.

Une dernière modalité de régulation, plus limitée, est celle qui consiste simplement à ne pas recommander certains contenus jugés "borderline" (ceux dont l'ambiguïté de la dimension insultante apparaît comme indécidable et ne peut être révélée ni par le constat d'une blessure ni par le calcul d'une probabilité de blessure). Ces contenus restent disponibles au niveau de l'archive de la plateforme mais ne sont pas activement suggérés par cette dernière. Cette stratégie est notamment de plus en plus adoptée par des plateformes comme YouTube ou Facebook lorsqu'elles cherchent à être le moins interventionnistes possible tout en répondant aux critiques selon lesquelles ne font pas qu'héberger des discours insultants mais les amplifient et les répètent à travers leurs logiques de recommandation (Gillepsie, 2022). Le double évitement recherché par cette dernière stratégie nous donne déjà une indication sur une dimension normative plus profonde de ces plateformes : non pas tant le fait qu'elles mettent à disposition des contenus (et potentiellement n'importe quel contenu) mais que cette mise à disposition intervient dans une économie de la rareté de l'attention où il s'agit de rendre certains contenus plus visibles que d'autres. En d'autres termes, on pourrait dire que la régulation des discours proposée par ces plateformes ne porte pas sur les discours en tant que tels mais sur les discours en tant qu'ils portent la marque de leur répétition et de leur efficacité algorithmique.

Les suites et recours au traitement algorithmique des discours de haine

Bien entendu, des recours restent possibles en interne : si un utilisateur estime que son contenu n'aurait pas dû être supprimé, il dispose de la possibilité de contester la décision qui sera alors réexaminée, réexamen dont les modalités ne sont pas précisées ("Je pense que Facebook n'aurait pas dû enlever ma publication.", *Facebook,* n.d. [2022]). Si l'utilisateur conteste la décision prise à la suite de ce réexamen, alors il pourra faire appel devant le Conseil de la Surveillance. Cet

appel ne peut être réalisé que sous certaines conditions, dont dépendent la sélection par le Conseil de l'appel ou son rejet, conditions dont l'utilisateur peut prendre connaissance en remplissant un questionnaire au début de la procédure sur le site du Conseil. Le Conseil explique par exemple avoir sélectionné des appels, par exemple à propos de "photos de la nudité pour accroître la sensibilisation aux symptômes du cancer du sein", ou encore à propos d'"une publication contenant une menace supposée pour avoir critiqué des convictions religieuses". Ce conseil est censé examiner de manière indépendante les décisions les plus délicates. Il est censé être composé de spécialistes internationaux issus d'horizons divers. Ses décisions ont par ailleurs un caractère contraignant puisqu'il peut annuler une décision préalablement prise, mais pas systématiquement puisqu'il peut aussi émettre des recommandations ("Oversight Board Bylaws", *Meta*, 2022). Il semble que cette procédure soit calquée sur le modèle des institutions juridiques, même s'il est entièrement internalisé et assoupli : pas de tiers, aucune garantie de traitement d'un recours assurée, absence de toute forme d'échange d'arguments, pas de publicité de la décision... Ces éléments sont assez évidents et ne sont pas l'objet premier de cet article. Cependant, est mise en exergue le fait qu'une telle voie *en apparence* plus judiciaire n'est en rien essentielle au développement des pratiques normatives qui nous occupent ici et qui se déroulent bien en amont.

3. Conséquences normatives depuis Derrida et Butler

Conflits de performatifs

Campons rapidement le paysage décrit ci-dessus à l'aide de quelques outils théoriques, à savoir depuis la théorie du performatif. Incontestablement, nous nous trouvons face à une série de conflits entre différents types d'énoncés performatifs : des discours de haine, dont le dire blesse, d'une part, et des normes, éventuellement organisées sous forme numérique, qui tentent de leur répondre. Les juristes se sont régulièrement référés à la théorie austinienne du performatif[2] pour légitimer une intervention règlementaire face aux discours blessants, a fortiori en contexte américain étant donné l'extrême méfiance qui y règne quant au fait de porter atteinte au principe de la liberté

[2] John L. Austin, *How to do Things with Words* nous invite à considérer les énoncés discursifs en ce qu'ils « font » quelque chose, en nous éloignant de l'illusion qui consisterait à réduire le langage à sa seule valeur descriptive, ou du moins à considérer que cette dernière nous en offre le sens premier. Les énoncés, loin de devoir être évalués seulement en termes de vrai et de faux, depuis leur valeur constative, en ce qu'ils disent le monde, ou encore en tant que propositions de nature « apophantique » selon le vocabulaire d'Aristote (*De l'interprétation*), peuvent aussi (et sans doute doivent toujours) être évalués en termes de réussi ou de raté, en ce qu'ils font quelque chose, en ce qu'ils participent à la fabrication du monde (Austin, 1962).

d'expression : un discours harcelant sur le plan sexuel (Catharine MacKinnon dans *Only Words* en 1993) ou des mots racialement connotés (les auteurs de la *Critical Race Theory* par exemple dans *Words that wound* en 1993) sont directement subordonnants, blessants ou menaçants pour celles ou ceux qui les reçoivent, ils peuvent ainsi être réfléchis en tant qu'actions, depuis ce qu'ils occasionnent directement à autrui, et non pas comme discours tels que protégés par la liberté d'expression.

Toutefois, en rebondissant pour une large part sur la critique derridienne de cette théorie du performatif, qui maintiendrait selon Derrida un reste d'idéalisme en ne pouvant concevoir l'échec d'un performatif que comme accidentel (avec en corolaire une fétichisation du contexte seul à même d'expliquer la réussite d'un performatif mais aussi une distinction nécessaire entre discours sérieux et discours non sérieux…) et en s'empêchant de saisir la nature profondément citationnelle de tout énoncé performatif (Derrida, 1972; Berns, 2018), Judith Butler nous fait remarquer non seulement que les injures peuvent échouer ou être détournées, mais surtout qu'il faut prendre acte du fait que leur réponse règlementaire peut les confirmer voire les instituer et les établir (Butler, 1997). Ce faisant, Butler nous indique la collusion qui se noue entre l'illusion d'une souveraineté du sujet (le sujet locuteur blesserait effectivement comme il le veut par les mots qu'il utilise mécaniquement - intention, énonciation et action coïncident parfaitement) et celle d'une souveraineté politique : l'État qui réglemente, sait ce qui blesse, veut en protéger ses sujets mais ainsi établit la blessure (Berns, 2021).

Pour Butler comme pour Derrida avant elle, le performatif doit se penser non seulement depuis une structure conventionnelle (entendue comme rapport intentionnel à un contexte), mais depuis sa structure itérable ou citationnelle, et donc aussi depuis sa possible décontextualisation, sa capacité à rompre par rapport à un contexte intentionnel et à subsister à cette rupture : une telle structure citationnelle désigne donc tout autant le caractère répétable, que l'altération que comprend toute répétition. Ou encore, à un niveau plus technique, il faut cesser de réduire le performatif à sa dimension illocutoire, avec l'incorporation de l'acte dans l'énoncé que ceci désigne, pour l'interpréter plutôt (a fortiori pour des actes de langage comme les injures) dans sa dimension perlocutoire (l'acte comme possible conséquence du dire, avec la distance que ceci installe entre le dire et la blessure, et donc la place que ceci laisse à la possibilité de voir des énoncés être détournés).

Gérer la viralité de manière indolore ?

Nous avons vu combien la gestion algorithmique des injures peinait à faire face à la multiplicité des contextes tout en restant entièrement soumise à l'idée que ceux-ci seuls déterminent la nature

d'une performance comme s'ils étaient disponibles ou devaient l'être. Nous avons vu aussi que ceci donnait lieu à une illusoire demande de clarification de ces contextes par l'explicitation des intentions des locuteurs (quand ils sont bien intentionnés !) mais aussi à un refus des plateformes d'assumer le fait qu'elles génèrent sans cesse des nouveaux contextes d'énonciation qui ne sont donc pas uniquement des expressions de contextes qui leur préexisteraient. Peut-être qu'une des raisons de ce déni est que ces "contextes" numériques nous mettent, plus que jamais, face à la dimension profondément citationnelle du langage : non seulement toute interaction numérique se présente et se légitime comme une citation du réel, mais une grande partie des interactions n'est explicitement rien d'autre que la citation d'un élément déjà présent sur le net, avec la dimension virale que nous connaissons désormais. En ce sens, de par ce mélange d'exigence de clarté et de déni, la réglementation algorithmique par les grandes plateformes risque non seulement de maintenir le langage dans sa dimension la plus illocutoire mais de l'exacerber en imaginant une scène linguistique parfaitement épurée où toute interaction pourrait se comprendre comme une mobilisation souveraine de sens.

Comment comprendre un tel retour à une approche très règlementaire du langage, avec la naïveté qui l'accompagne (réclamer la clarification des intentions, ne pas percevoir la prolifération des contextes...) ? Nous devons ici nous arrêter sur deux aspects, profondément corrélés entre eux, spécifiques à la dimension numérique du phénomène du langage, et par dimension numérique il faut entendre le fait que le cadre de l'injure aussi bien que la réponse qui lui est apportée sont de nature numérique. Il s'agit d'une part de vouloir répondre non pas tant aux injures elles-mêmes qu'à leur viralité, à leur répétition, avec la particularité que celle-ci se présente sous la forme suivante : certes cette viralité est empreinte de variation, de changement de contexte mais l'injure est considérée en son sein comme si elle se maintenait intacte au sein-même de ces mouvements. La perspective derridienne de la citation semble de ce point de vue exactement renversée : au sein du changement et de la variation, le même, l'identité à soi de l'injure, continue d'être ce à l'aune de quoi toute citation sera observée. Il s'agit d'autre part, et en conséquence de vouloir faire face à ceci de la manière la plus radicalement indolore, en s'approchant le plus possible d'un effacement de l'injure, à la racine, à son origine même. A nouveau, la perspective derridienne de la citation semble ainsi exactement renversée dans cette prétention à saisir un phénomène citationnel au plus près de son origine, sans la différer.

Si on se réfère maintenant, sur une telle base, aux hypothèses butleriennes selon lesquelles la réponse règlementaire pourrait être une manière d'établir voire de décréter la blessure, la réponse algorithmique aurait pour particularité d'instaurer les déplacements suivants. D'une part, un tel établissement de la blessure pourrait (sembler) être évité puisqu'il serait, idéalement, possible

d'effacer comme tel l'insulte, de ne pas la laisser apparaître (même si on a vu combien c'est là un fantasme). D'autre part, et en conséquence, est semblablement empêchée toute forme de déplacement, de reprise décontextualisante de l'injure, puisque tout simplement celle-ci n'apparaît pas… et que toute forme de reprise est d'abord suspecte, sauf accompagnée d'une clarification de la bonne intention du locuteur !

Bien sûr, il s'agit là d'une scène idéale ! Dans les faits, la régulation algorithmique des injures ne permet pas d'éviter la violence inhérente à toute réglementation. Il ne s'agit pas là d'un élément de critique facile mais d'un constat réaliste : toute régulation porte un fond de violence incompressible, établit des partages entre ce qui peut être dit et ce qui ne peut pas être dit, entre ce qui est légitime ou illégitime, qui sont institués comme tels par l'institution qui a en charge cette régulation. Or, et c'est sans doute là la grande différence entre la réponse "pro-active" ou automatique proposée par les plateformes et la réponse plus classiquement juridique, la première ne se pense pas comme une institution qui aurait pour responsabilité de tels partages, et dans un tel cadre, la blessure ne comparaît d'aucune manière, elle est évitée. Sans même parler de la difficulté à accéder à la convention, ou au code informatique à partir duquel l'interdiction est opérée, la dimension secrète et virtuelle de l'archive pointée ci-dessus dans les cas de suppression de contenus, suffit à souligner cet évitement. Là où le droit doit citer l'injure pour pouvoir l'interdire - la confirmant de la sorte mais montrant déjà, en creux, par cette citation même, que le déplacement de sens et des effets est possible – le réglementation de plus en plus automatisée doit de moins en moins passer par une telle citation et ce a fortiori avec des systèmes qui reposent sur un apprentissage par des algorithmes de ce qui *risque* de blesser, apprentissage nourri par des discours passés. Plus encore, la citation, et ce qu'elle implique en termes d'épreuve dont la réussite est toujours fragile ou incertaine, se dissout dans le processus d'"apprentissage" algorithmique dont la spontanéité et la naturalité apparentes seraient étrangères à toute artificialité qui marque l'exercice du pouvoir juridico-politique classique. Par cette rapide confrontation entre régulation par le droit et régulation automatisée, nous ne voulons pas plaider pour le maintien, utopique, du monopole de la première, mais simplement pointer, hors de toute considération quant au fait de savoir quels sont les agents normatifs légitimes, ce qui est au final dissout dans la seconde, à savoir: l'épreuve même de la citation, avec l'incertitude qu'elle comprend et dont elle témoigne.

L'évitement de toute blessure, de toute violence par les grandes plateformes numériques frappe d'autant plus par l'incroyable ubiquité et ampleur de leurs médiations. Aucun système de régulation juridique des discours ne peut se targuer d'une telle extension planétaire, ni d'une telle granularité de son insertion quotidienne. L'ampleur du système de citation numérique ne repose pas simplement sur un oubli des particularités et des contextes locaux à partir desquels on pourrait

reconstruire le sens “véritable” (et donc les effets “réels”) en cas de différend ou d’indétermination, mais constitue le nouveau contexte d’énonciation sur une scène linguistique désormais profondément virale et citationnelle. Les performances des systèmes apprenants dont les sorties sont produites en boucle avec leurs “utilisateurs” humains dans une logique probabiliste et non plus déterministe, nous éloignent eux-mêmes d’une répétition machinique qui serait purement illocutoire et nous rapprochent d’une efficacité perlocutoire marquée par l’indisponibilité de la convention, du contexte et des locuteurs d’”origine” (Berns et Reigeluth 2021 pp. 127-131; Reigeluth 2023). La machine n’est pas condamnée à être l’illustration d’un illocutoire idéalisé, son efficacité se présente à travers son ouverture à des interactions partiellement indéterminées. Disant cela, nous ne voulons pas soutenir que la dimension numérique du langage serait radicalement autre, mais au contraire qu’elle nous confronte plus que jamais à ce qu’une conception idéalisée du langage aurait rendu incompréhensible, à savoir le fait que le langage produit toujours des effets au-delà de son contexte et de son intention. Reconnaître ce débordement du langage sur lui-même n’implique pas forcément un laisser-dire ou un abandon de toute tentative de réguler les discours, mais nous enjoint à reconnaître que l’interdiction, qu’elle soit algorithmique ou juridique, ex ante ou ex post, ne peut manquer de relancer la machine citationnelle, c’est-à-dire l’usage effectif du langage. Ce que le traitement algorithmique nous invite à penser, au minimum, c’est une régulation du langage pour lequel il n’y aurait plus de dehors, et qui donc participe elle-même pleinement au langage.

Nous aimerions clore en indiquant un enjeu éthique et politique ultime qui demanderait à être exploré plus en avant et qui ouvre sur une articulation plus profonde entre le traitement algorithmique des discours de haine et la liberté d‘expression. L’interdiction ne constitue pas le dehors ou la limite du discours, mais son intériorité mouvante. C’est *parce que* les mots risquent toujours de blesser que nous pouvons encore faire attention aux mots que nous utilisons, que certains mots ont plus de poids que d’autres, que nous pouvons discerner leurs différents effets possibles et être étonnés par leurs effets ou reprises imprévus. Le danger d’une régulation automatique des discours – et nous ne visons pas ici une tendance inhérente au “numérique” mais bien les rapports de pouvoir qui se jouent sur ses plateformes plus monopolistiques – qui chercherait à éviter la blessure avant même qu‘elle ne survienne réside dans l’illusion de pouvoir libérer les sujets du poids des mots, aussi bien pour ceux qui subiraient une injure que pour ceux qui seraient susceptibles de blesser et qui ne “peuvent plus rien dire”. Ce qui ne revient pas à dire que cela nous priverait simplement d‘une occasion d’exercer nos responsabilités individuelles mais surtout d’une épreuve de la conflictualité de la vie sociale qui s‘exprime dans le langage. En d’autres termes, la question n’apparaît absorbable ni par la seule réflexion éthique, ni par une approche purement juridique mais renvoie toujours aux rapports de force politiques au sein desquels le dicible et l’indicible se partagent. Le fait de toujours pouvoir dire “tu ne

peux pas dire ça", quand bien même aucune sanction formelle ne le soutient, ou de toujours pouvoir dire ce qu'on nous a interdit de dire, - sorte de réserve de droit naturel au sens spinoziste qui témoigne de l'obstacle incompressible auquel toute prétention de souveraineté absolue doit faire face en gouvernant les actes et les paroles des sujets (Spinoza 1965 pp. 277-279) – indique bien que la limite du langage ne cesse de se jouer au travers de son usage même. L''idéal d'un langage qui s'autorégulerait par un processus purement technique (mais dont l'efficacité reposerait sur son refus de toute artificialité) risque non seulement de rater cet usage effectif mais de rendre inconséquent la possibilité même d'être blessé.

Références :

Matériau de recherche :

"Discours haineux.", *Meta Transparency Center*, (n.d.), [consulté le 17/12/2022], URL : https://transparency.fb.com/fr-fr/policies/community-standards/hate-speech/.

"Je pense que Facebook n'aurait pas dû enlever ma publication.", *Facebook,* [consulté le 17/12/2022], URL : https://www.facebook.com/help/2090856331203011?helpref=faq_content.

"Standards de la communauté Facebook", *Meta Transparecy Center*, (n.d.), [consulté le 17/12/2022], URL : https://transparency.fb.com/fr-fr/policies/community-standards/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2F.

"Oversight Board Bylaws", *Meta*, Janvier 2022, [consulté le 17/12/2022], URL : https://about.fb.com/wp-content/uploads/2020/01/Bylaws_v6.pdf.

Ouvrages :

Austin, J.L., *How to do things with words*, Oxford University Press, 1962.

Berns, T., "De la gravité de la loi au prosaïsme du droit, avec Derrida", *Ethique, politique, religions*, n°12, 2018-1, p.45-58, DOI : 10.15122/isbn.978-2-406-08298-9.p.0045.

Berns, T., "Insult and Post-sovereign Law as Juridicity", *Political theology*, vol.22, n°2, 2021, p.147-154, DOI : https://doi.org/10.1080/1462317X.2021.1885828.

Berns, T., Reigeluth, T., *Ethique de la communication et de l'information. Une initiation philosophique et situation technologique avancée*, Presses Universitaires de Bruxelles, 2021.

Butler, J., *Excitable Speech. A Politics of the Performative*, Routledge, 1997.

Derrida, J., "Signature, événement, contexte : Écriture et télécommunication", *Marges de la philosophie*, Les Éditions de Minuit, 1972.

Gillepsie, T., *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*, Yale University Press, New Haven, 2018.

Gillepsie, T., "Do Not Recommend ? Reduction as a Form of Content Moderation.", *Social Media + Society*, vol.8, n°3, 2022, DOI : https://doi.org/10.1177/20563051221117552.

Gorwa, R., Binns, R., Katzenbach, C., "Algorithmic content moderation : Technical and political challenges in the automation of platform governance", *Big Data & Society*, vol.7, n°1, 2020, DOI : https://doi.org/10.1177/2053951719897945.

Herwanto, H.G., Trisna, N.P., "Hate Speech and Abusive Language Classification Using fastTest", *International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 2019, DOI : https://doi.org/10.1109/ISRITI48646.2019.9034560.

Martins, R., Gomes, M., Almeida, J.J., Novais, P., Henriques, P., "Hate speech classification in social media using emotional analysis", *7th Brazilian Conference on Intelligent Systems (BRACIS),* 2018, DOI : http://dx.doi.org/10.1109/BRACIS.2018.00019.

Putri, T.T.A., Sriadhi, S., Sari, R.D., Rahmadani, R., Hutahaean, H.D., "A comparison of classification algorithms for hate speech detection", *IOP Conference Series : Materials Science and Engineering,* 2020, DOI : http://dx.doi.org/10.1088/1757-899X/830/3/032006.

Reigeluth, T., « Machine Learning Normativity as Performativity », Lindgren, S., (éd.) Elgar E., *Handbook of Critical Studies of Artificial Intelligence*, 2023 (à paraître).

Spinoza, *Traité théologico-politique*, trad. Ch. Appuhn, Œuvres II, Paris, Garnier-Flammarion, 1965.

Vinot, R., Grabar, N., Valette, M., "Application d'algorithmes de classification automatique pour la détection des contenus racistes sur l'Internet", *Acte de la 10ème conférence sur le Traitement Automatique es Langues Naturelles. Articles longs*, 2003, p.275-284, URL : https://aclanthology.org/2003.jeptalnrecital-long.26.