Le casse-tête de la modération des réseaux sociaux

par Ophélie Surcouf

Depuis un peu plus d'un an, les grandes entreprises de la Silicon Valley tentent de mieux réguler leurs plateformes face à la pression des gouvernements et la colère des utilisateurs du monde entier. Haine, désinformation, incitations à la violence, harcèlement, manipulation: comment endiguer ces comportements qui gangrènent les réseaux utilisés au quotidien par des milliards d'individus? Différentes pistes existent pour permettre des échanges sereins tout en préservant la liberté d'expression, enjeu cardinal de ce dilemme.

Pourquoi c'est important. La modération affecte tous les contenus en ligne et, donc, chacune de nos interactions avec le web. Pour le moment, chaque site ou entreprise décide de ses règles pour déterminer ce qui est acceptable et ce qui ne l'est pas - qu'il s'agisse des premiers résultats de Google, des posts Facebook, Instagram ou Twitter, des commentaires sur Youtube, des conversations sur Whatsapp, Messenger ou Telegram, des forums comme Reddit ou d'une page Wikipedia. Mais cette technique du cas par cas montre ses limites depuis des années: les réseaux sociaux et les applications de messagerie voient les discours haineux proliférer dans les fils de leurs utilisateurs et ne parviennent pas à contenir la déferlante.

Les racines du problème. Selon Lea Strohm, lab-manager d'ethix, un laboratoire Suisse d'éthique de l'innovation :

- **Un conflit d'intérêts fondamental pour les entreprises**. Celles-ci ne peuvent pas prioriser leur croissance et faire une modération efficace. «Facebook, par exemple, a toujours jugé que l'interaction entre ses utilisateurs était plus importante parce que c'est la métrique qu'ils utilisent pour vendre la publicité ou convaincre les investisseurs.»
- **La faillibilité des algorithmes.** Leurs mécanismes de sécurité peuvent être facilement contournés (en remplaçant des lettres par des chiffres ou des espaces par exemple) et ils font souvent des erreurs lorsqu'ils s'agit d'analyser les contenus (parmi les drames récents, Facebook a censuré une photo de la ville de Bitche dans le sud de la France). «Leurs performances sont de plus inégales en fonction des langues», souligne Lea Strohm. Les langues les plus parlées sur le web, comme l'anglais, ont davantage de contenus - les algorithmes sont donc mieux entraînés à les repérer et les analyser. Cela entraîne des temps de latence plus longs lorsqu'il s'agit de modérer les contenus et quelques secondes peuvent s'avérer critiques sur internet…
- **Pour les modérateurs, des conditions de travail préoccupantes**, notamment sur le plan psychologique, car les employés sont quotidiennement exposés à des contenus très sérieux ou violents.

L'avis du modérateur. Patrick O'Keefe modère des communautés depuis plus de 20 ans. Il est l'auteur de Managing Online Forums, anime le podcast Community Signal sur la modération et les innovations dans le milieu. Il est également consultant pour CNN sur les communautés en ligne.

> «Lorsque l'on parle des limites de la modération, je pense que l'on a surtout Facebook, Twitter et autres grands réseaux sociaux en tête. Mais c'est oublier que 99,9% des communautés en ligne sont plus petites et, qu'en général, elles

> ont réglé ce genre de problèmes avec un certain degré de réussite: la plupart ne croulent pas sous les messages nazis. Nous sommes ce que nous tolérons. Ce qui pose la question: à quel moment devient-on trop gros pour proposer un environnement sain? Les gens supposent que les grands réseaux sociaux doivent exister, mais est-ce vraiment le cas s'ils sont trop grands pour être gérés?»

Les solutions envisagées.

- **Mieux labelliser les contenus pour offrir un meilleur contexte à l'utilisateur.** C'est une piste explorée par Facebook pour mieux indiquer d'où vient l'information (jusque-là, il fallait installer des plug-ins sur les moteurs de recherche pour obtenir ce genre d'avertissements). Le géant américain a annoncé tester cette fonction qui consistera à indiquer «fan page», «page publique officielle» ou encore «média satirique» sur des posts apparaissant dans le fil de l'utilisateur. Le risque est toutefois que les utilisateurs malveillants, y compris les extrémistes violents, puissent utiliser ce moyen pour cacher du contenu et réduire les chances que les utilisateurs le signalent. Tech Against Terrorism, une initiative soutenue par les Nations unies pour aider l'industrie technologique à lutter contre l'utilisation d'Internet à des fins terroristes, a déjà signalé à des plateformes des contenus étiquetés par l'utilisateur qui les mettait en ligne comme «NSFW» (Not Safe For Work). L'organisation recommande donc de prêter attention aux contenus étiquetés par l'utilisateur comme «NSFW» (ou tout autre étiquette/filtre pouvant être appliqué).
- **Décentraliser**. C'est-à-dire laisser à des tiers la responsabilité de gérer la modération d'une plateforme, voire à l'utilisateur (comme l'étrange logiciel d'Intel, Bleep, qui permet aux gamers de filtrer le contenu audio en fonction de leurs préférences). «Je ne veux pas décourager les outils qui permettent aux gens de prendre davantage de contrôle de leur fil d'actualité», constate le modérateur Patrick O'Keefe. Par exemple en pouvant filtrer les résultats de ses recherches, censurer certains mots, bloquer ou mettre en sourdine d'autres utilisateurs… «Nous avons tous la responsabilité de protéger notre propre bien-être et il est important d'en avoir les moyens», ajoute-t-il. Cela dit, cette possibilité ne devrait pas dédouaner les plateformes de leurs responsabilités vis-à-vis de leurs utilisateurs. Ou leur permettre de sous-payer un personnel peu qualifié pour limiter les dépenses.
- **Élargir le champ d'action aux faits qui se produisent en dehors de la plateforme.** Twitch vient d'étendre ses règles de modération: désormais, si un livestreamer de Twitch est harcelé en dehors de Twitch (par exemple Twitter), Twitch permet à un cabinet d'avocat tiers d'enquêter et d'agir. Une première qui doit encore faire ses preuves dans une industrie qui jusque-là restait très territoriale sur ce genre de cas.
- **Plus de transparence auprès des internautes modérés**. «La plupart des améliorations vient des petites choses, pas de grands algorithmes ou d'intelligences artificielles», défend Patrick O'Keefe. «Par exemple, une idée toute bête que j'ai développée dans une de mes communautés: lorsqu'un mot bloqué active le programme que nous avons développé, l'utilisateur reçoit un message lui disant "voici le mot qui a provoqué le blocage du message, si vous voulez qu'il soit publié, vous devez ajuster votre propos". Tout le monde est content: nous n'avons pas de mots inappropriés, l'utilisateur peut poster son commentaire et nous n'avons jamais à le retirer! Cela peut sembler élémentaire et cela ne s'applique pas à tous les espaces - mais de petites choses peuvent avoir un impact important.»
- **Publier des rapports avec des chiffres et des analyses.** «Nous conseillons aux entreprises de contextualiser leurs résultats lorsqu'elles font leurs rapports sur la modération», explique Maygane Janin, analyste chez Tech against terrorism. «Que leur transparence ait un sens et ne soit pas une

énumération de chiffres qui n'évoque rien à l'utilisateur. Discord fait beaucoup d'efforts dans ce sens. Facebook et Youtube ont de leur côté fait des progrès, la crise sanitaire ayant beaucoup changé la donne. Tik Tok inclut aussi des retours sur ce qu'il a essayé d'implémenter au cours de l'année écoulée.» Le problème est que, si les grandes entreprises ont les moyens de transmettre ces conclusions au grand public, les petites entreprises ne peuvent pas fournir le même niveau d'information et de détail du fait d'un manque de ressource… «Les différents rapports devraient refléter chaque les différentes politiques et pratiques de modération qui existent», ajoute Maygane Janin. «Cette diversité doit être encouragée pour préserver un internet diversifié et dynamique.»

- **Changer les mécanismes des réseaux.** «Certains prétendent que le problème vient des publicités ou des flux algorithmiques (deux idées avec lesquelles je ne suis pas du tout d'accord). Cela dit, ils soulèvent le même point sous-jacent: au lieu de chercher les mauvais éléments, peut-être devrions-nous changer les chemins que les mauvais éléments peuvent emprunter», écrivait Ben Evans dans un essai intitulé Est-ce que la modération de contenu est dans une impasse?. En clair, et si on changeait ce que l'on peut faire sur les réseaux, plutôt que de l'encadrer? Pinterest a publié une charte pour ses créateurs afin de doter ceux-ci de plus de pouvoir et de donner le ton d'emblée à l'ambiance de sa plateforme. Instagram vient de se lancer dans une expérience: cacher les comptes de likes sur certaines publications. Twitter demande depuis récemment si on a lu le contenu du lien que l'on s'apprête à retweeter afin de limiter la viralité des tweets. Quant à retirer l'anonymat (l'une des raisons les plus évidentes pour de nombreux cas de harcèlement en ligne), l'idée fait des allers-retours depuis des années entre les gouvernements, les plateformes et les régulateurs - sans toutefois être actée.

L'avis du mathématicien. Celui de Paul-Olivier Dehaye, fondateur de PersonalData.IO et membre du conseil d'administration de MyData Global, est l'expert à l'origine de l'affaire Cambridge Analytica:

> «Le problème ne réside pas dans la modération mais provient du fait qu'on essaie de créer une agora la plus large possible sur la base d'un mauvais modèle d'affaires. Personne n'a la prétention d'aller dans un train et de dire tout haut "ayons un débat sur l'avortement, ici et maintenant, cela va bien se passer". D'autant plus que Mark Zuckerberg propose quelque chose d'encore plus fou: le débat anonyme de deux personnes assises dans des trains à des milliers de kilomètres de distance. Cela ne peut pas fonctionner.»

Et d'évoquer une autre voie:

> «Une proposition plus réaliste, quoique imparfaite, émane de Reddit. Ce site permet de construire des communautés plus petites qui se réunissent autour d'un centre d'intérêt particulier. Elles sont modérées par leurs membres et Reddit les dote d'outils pour cela. Si quelqu'un a un mauvais comportement, le modérateur n'a pas à se poser beaucoup de questions et peut éjecter la personne du groupe. Ce n'est pas un bannissement permanent d'une vie numérique. Alors que sur Facebook, si vous êtes banni, vous êtes complètement exclu d'un service très transversal.»

Qui doit décider? Dans les faits, les réseaux sociaux se réservent le droit de déterminer ce qui est vrai et ce qui est faux. Mais est-ce leur rôle? Pour donner une chance à la modération, il est essentiel de s'accorder sur des règles.

Une mission prise à bras le corps par Facebook qui est, avec Twitter, le réseau le plus pris dans le collimateur des régulateurs. Depuis octobre 2020, son Conseil de surveillance délibère sur des cas sensibles de modération des contenus sur la plateforme pour créer une jurisprudence. Une démarche qui fait tiquer de nombreux observateurs qui reprochent aux membres du Conseil de n'être que des marionnettes politiques de l'entreprise et s'attribuer des pouvoirs en dehors de ses attributions.

L'avis de l'avocat. Yaniv Benhamou est spécialisé dans la propriété intellectuelle, la protection des données et chargé de cours à l'Université de Genève. Avec le Digital Law Center de la Faculté de droit, il travaille sur des projets de recommandations politiques dont le but est l'émergence de standards internationaux pour la résolution de litiges sur Internet.

Yaniv Benhamou:

> «Facebook prend un pari risqué avec ce Conseil, mais cela pourrait fonctionner. Ses décisions pourraient être imitées par d'autres plateformes et pousser le droit du numérique à converger et à se globaliser.»

L'universitaire relève que les obstacles juridiques proviennent surtout de la difficile mise en œuvre des droits. De fait, les réseaux sociaux ont des effets globaux, tandis que le droit demeure local et fragmenté. L'origine des contenus incriminés est souvent difficile à déterminer ou localisée à l'étranger, ce qui suppose de recourir à des procédures complexes d'entraide internationale. Quant à l'encadrement des réseaux sociaux, il reste encore timide même si, après une régulation souple (par exemple, avec l'auto-régulation), on semble glisser progressivement vers une régulation forte. L'Union européenne est en train de travailler à renforcer l'encadrement et la surveillance des plateformes (on peut citer les lois sur les marchés et services numériques).

Yaniv Benhamou:

> «Parmi les obstacles pratiques: il est difficile de définir ce qu'est une information acceptable. Et en admettant que l'on y parvienne, par exemple grâce à un code de conduite universel ou une définition globale, les plateformes sont noyées dans leur propre masse d'informations. Des informations qui se situent souvent en zone grise et sont ainsi difficiles à identifier et à interpréter par des algorithmes. Sera-t-il seulement possible de programmer un algorithme techniquement capable de reconnaître des propos haineux?»