Discriminations algorithmiques, modération des réseaux sociaux et militantisme

Les entreprises détentrices des réseaux sociaux comme Twitter, Facebook, YouTube, TikTok (pour ne citer que ceux-là) ont de plus en plus recours à des systèmes d'IA pour réguler les contenus en ligne. Cependant, cette délégation de la modération à des outils automatisés semble aller de pair avec l'apparition de nouvelles formes de discriminations en ligne. Ainsi, en plus de faire face aux violences et attaques de groupuscules d'extrême droite, les militant·e·s de gauche et issu·e·s de communautés minorisées sont de plus en plus confronté·e·s à des risques de censure abusive en ligne. Thibaut Grison explique dans cet article pourquoi se saisir des technologies algorithmiques constitue un enjeu politique en particulier pour la gauche.

En mai 2022, l'éventuel rachat de Twitter par Elon Musk a fait prendre conscience du caractère éminemment politique et marketing de la modération d'un réseau social. Par ces annonces, le milliardaire propriétaire de SpaceX et de Tesla a fait savoir qu'il comptait « œuvrer pour une plus grande liberté d'expression sur le réseau social », en réduisant le champ d'action des processus de modération - chargés, entre autres, de réguler les contenus haineux, pédopornographiques, faisant l'apologie du terrorisme ou encore les *fake news* - et en basculant les algorithmes de recommandation et de modération de l'oiseau bleu en « open source ». Alors que de nombreux·ses militant·e·s LGBT+, féministes et antiracistes ont craint une recrudescence de cyber-harcèlement à leur encontre et que certain·e·s ont même fait le choix de boycotter le réseau social, les militants conservateurs et d'extrême droite ont accueilli la nouvelle à bras ouverts en y voyant une revanche sur la fermeture du compte de Donald Trump à la suite des évènements du Capitole. En effet, l'enjeu de la modération des réseaux sociaux a fait l'objet de nombreuses polémiques ces dernières années ; en particulier, en ce qu'elle favoriserait l'émergence de certains discours fascistes et réduirait au silence l'expression de certaines identités minorisées.

Avoir recours à l'IA pour modérer les réseaux sociaux : danger social ou solution technique ?

La modération des réseaux sociaux désigne le processus, les stratégies et les textes qui encadrent la visibilité des contenus et profils des différentes plateformes en ligne. Si les premiers forums de discussion sur internet avaient déjà recours à de la main-d'œuvre humaine pour encadrer la bonne conduite des échanges en ligne, c'est l'arrivée des plus gros réseaux sociaux comme Facebook, Twitter ou YouTube qui a conduit les entreprises détentrices de ces services à professionnaliser le travail de modération par l'emploi d'une main-d'œuvre dédiée.

Cette main-d'œuvre est en réalité souvent rémunérée par des entreprises sous-traitantes démarchées par les GAFAM. Disons-le sans détour, les travailleurs et travailleuses du clic chargé·e·s de maintenir la bonne tenue des échanges en ligne subissent des conditions de travail déplorables : sous-staffé·e·s, confronté·e·s quotidiennement à des contenus pédopornographiques, terroristes ou haineux (entre autres contenus traumatisants) et sous-payé·e·s ; cette force de travail est soumise à de lourdes pressions psychologiques en plus d'être totalement invisibilisée. En effet, le personnel de modération exerce dans des conditions totalement opaques : importantes clauses de confidentialité sur la nature et l'ampleur des contenus modérés, incertitude sur le nombre de modérateur·rice·s employé·e·s par chaque entreprise, etc[1]. Rappelons que ce travail de modération s'appuie également sur une main-d'œuvre totalement bénévole qui, quotidiennement, signale les contenus illicites auprès de la plateforme. Ce travail bénévole est en partie opéré par des militant·e·s issues de communautés minorisées qui, chaque jour, signalent auprès des plateformes les contenus dont iels sont victimes. La modération de contenus s'apparente donc à un travail numérique précaire propre au système économique capitaliste et néolibéral dans lequel nous vivons.

Dans ce contexte, l'investissement dans des algorithmes et systèmes d'IA pour réguler la diffusion de contenus illicites postés sur les réseaux sociaux apparaîtrait comme une alternative technologique et sociale opportune. En effet, la prolifération de contenus haineux depuis la succession des périodes de confinement, les polémiques sur les conditions de travail des modérateur·rice·s de contenus ou encore les injonctions des pouvoirs publics à modérer plus efficacement ce qui relèverait de la désinformation, poussent les entreprises détentrices des réseaux sociaux à recourir à des systèmes automatisés pour réguler la liberté d'expression en ligne. Mais, ce recours à des algorithmes entraîne paradoxalement une cristallisation des systèmes de domination en ligne…

Algorithmes de modération et discriminations algorithmiques

Le vote de la loi du 24 juin 2020 proposée par la majorité La République en Marche et visant à « lutter contre les contenus haineux sur internet » - dite « loi Avia » - a fait craindre à certains groupes de l'opposition parlementaire et à des militant·e·s LGBT+, antiracistes et féministes un recours accru aux algorithmes pour modérer les contenus jugés illicites en ligne. En effet, iels ont estimé que le délai maximum de 24h imposé dans cette nouvelle loi aux entreprises détentrices des réseaux sociaux (comme Twitter, Facebook, YouTube, etc.) allait encourager la modération exclusivement automatisée. Et pour cause, en mai 2020, alors que la proposition de loi est encore en débat sur les bancs de l'Assemblée nationale, une vingtaine de comptes de militant·e·s LGBT+ sont suspendus du jour au lendemain de Twitter et Facebook. Les militant·e·s font alors l'hypothèse que des algorithmes de modération n'auraient pas fait la distinction entre la réappropriation militante de termes injurieux comme « pédé » et « gouine » et des discours de haine en ligne et auraient donc supprimé ces profils des plateformes. Si la disposition du délai de 24h a été censurée par le Conseil constitutionnel car elle présentait un risque démesuré pour la liberté d'expression des internautes ; les cas de discriminations algorithmiques dans le contexte de la modération des réseaux sociaux demeurent.

Sur TikTok, Instagram et YouTube, un certain nombre de créateur·rice·s ont remarqué que des contenus pouvaient être invisibilisés par les dispositifs de modération automatisée dès lors qu'ils faisaient mention d'enjeux d'identité sexuelle et de genre ou liés à la race. On appelle cette forme d'invisibilisation du « shadowbanning ». Il s'agit d'une forme de censure insidieuse consistant en un déréférencement du contenu des fils d'actualités, *timelines* et pages de recommandation des internautes sans que le contenu ne soit supprimé du profil des internautes. En d'autres termes, les créateur·rice·s de contenus sont censurés sans même qu'iels ne puissent s'en apercevoir. Selon une enquête menée par le think tank australien ASPI en 2020[2], cette stratégie de censure aurait notamment été employée par TikTok pour censurer les vidéos référencées sous le hashtag #BlackLivesMatter aux États-Unis afin que les militant·e·s antiracistes ne puissent pas s'en rendre compte et les empêchent de créer les conditions d'une polémique qui nuirait à l'image de marque de l'entreprise Bytedance détentrice de TikTok. Ainsi, la discrimination algorithmique sur les réseaux sociaux s'apparente à une forme de censure dont les internautes n'ont parfois même pas conscience d'être victimes.

Les causes de censure en ligne à l'encontre des militant·e·s sont multiples. Elles sont parfois dues à des signalements abusifs de groupuscules d'extrême droite qui se coordonnent pour causer la fermeture d'un compte en particulier. Elles peuvent également être le fait de décisions politiques des propriétaires des plateformes. Mais, elles semblent le plus souvent être causées par des biais algorithmiques. En effet, la qualité et la nature des données - ou encore la définition d'impératifs de rentabilité économique - à partir desquels les algorithmes de modération et de recommandation sont entraînés, peut conduire à des cas de discriminations algorithmiques sur les réseaux sociaux numériques. L'implémentation de certains biais dans le déploiement de ces algorithmes peut également favoriser la prolifération de contenus émanant d'internautes d'extrême droite en raison du fort taux d'interaction qu'ils suscitent. Dans un billet publié sur leur blog officiel en octobre 2021[3], Twitter a notamment admis que « les tweets publiés par des comptes de la droite reçoivent plus d'amplification algorithmique que la gauche », sans qu'ils en connaissent les raisons… Les entreprises détentrices des réseaux sociaux peuvent donc favoriser une orientation politique par rapport à une autre. Et, en l'occurrence, les militant·e·s de gauche, féministes, LGBT+ et antiracistes semblent être les principales victimes de cette censure algorithmique… Édifiant !

Militantisme et réseaux sociaux : se saisir des algorithmes !

Bien que les activistes de gauche aient su se saisir des réseaux sociaux afin de faire émerger des mouvements sociaux de grande ampleur (*MeToo*, *BlackLivesMatter* ou encore le mouvement des *Indignados* en Espagne et des Gilets jaunes en France) ; l'expression publique de certaines paroles militantes semble encore menacée. L'enjeu des discriminations algorithmiques nous rappelle que les réseaux sociaux tels que Twitter, Facebook, YouTube ou TikTok sont avant tout des entreprises concentrées entre les mains de quelques milliardaires qui suivent leurs propres intérêts économiques. Avant d'être des espaces de liberté d'expression et de libération de la parole militante, les réseaux sociaux sont des entreprises marchandes qui éditorialisent les contenus, mettent en avant certaines idées et courants politiques et invisibilisent certains propos indésirables selon leurs propres termes et aux moyens d'outils technologiques opaques. Autrement dit, bien que les chartes de modération et communiqués de presse officiels des entreprises capitalisent sur la promotion de valeurs multiculturelles, telles que le respect de la diversité des internautes, l'application de ces chartes de modération n'en est pas moins discriminante.

Par ailleurs, il existe un certain paradoxe dans la posture des États et des puissances publiques qui disent vouloir réguler l'impact des GAFAM ou prôner davantage de lutte contre la prolifération de contenus haineux en ligne, tout en délégant le travail de censure à des algorithmes et plus largement à des entreprises privées. Et ce, sans mettre en place tout un ensemble de garde-fous qui permettraient de réguler la façon dont ces entreprises nous régulent, avant de leur donner plus de prérogatives. En effet, les algorithmes mobilisés dans la modération et la recommandation des contenus des réseaux sociaux présentent un nouveau risque pour le cyber-activisme. Alors que les sujets discutés sur les réseaux sociaux influencent aujourd'hui les sujets portés par les médias traditionnels, les dispositifs de modération encadrent de plus en plus les périmètres de la contestation en ligne. Se saisir des technologies algorithmiques apparaît donc comme un véritable enjeu du militantisme de gauche aujourd'hui, au risque de se voir dessaisir de certains droits fondamentaux comme celui de la liberté d'expression et de contestation en ligne.

Notes de bas de page :

[1] Antonio Casili, *En attendant les robots – enquête sur le travail du clic*, Seuil, Paris,2021.

[2] Fergus Ryan, Audrey Fritz et Daria Impiombato, «Tiktok and WeChat: Curating and controlling global information flows», ASPI, International Cyber Policy Centre, 2020.

[3] Rumman Chowdhury et Luca Belli, «Examining algorithmic amplification of political content on Twitter», blog de Twitter, le 21 octobre 2021.

Pour citer cet article :

Thibault Grison, «Discriminations algorithmiques, modération des réseaux sociaux et militantisme», *Silomag*, n°15, juillet 2022. URL: https://silogora.org/discriminations-algorithmiques-moderation-des-reseaux-sociaux-et-militantisme