L'Europe veut encadrer les algorithmes pour retirer les contenus illicites et éviter les " faux positifs "

Winston Maxwell

L'Europe veut encadrer les algorithmes pour retirer les contenus illicites et éviter les « *faux positifs* »

Le futur règlement européen Digital Services Act (DSA) veut encadrer l'utilisation d'algorithmes dans la gestion des contenus sur les réseaux sociaux et d'en retirer ceux « jugés » illicites. Mais le risque de « *faux positifs* » (bloqués à tort) va poser des problèmes aux régulateurs et aux juges.

Par Winston Maxwell*, Telecom Paris, Institut polytechnique de Paris

Bloquer la publication d'un contenu est une décision grave, portant potentiellement atteinte à l'un des droits fondamentaux les plus importants pour la démocratie : la liberté d'expression. Pour la préserver, le droit constitutionnel américain et français exigent généralement qu'une décision interdisant la diffusion de contenus soit prise par une autorité judiciaire, et qu'elle le soit prise après la publication du contenu, non avant (1).

Blocage automatique : quelle légitimité ?

Les plateformes ne s'embarrassent pas de ces principes, filtrant des contenus avant leur publication par l'utilisation de robots. Faut-il s'en inquiéter ? S'agit-il d'une violation des droits fondamentaux des utilisateurs ? Le recours aux algorithmes pour identifier des contenus illégaux est devenu incontournable en raison de la quantité des informations publiées par les utilisateurs des réseaux sociaux. Même si la loi n'impose pas aux plateformes une obligation générale de surveillance des contenus, laquelle reste interdite (2), celles-ci ont mis en place des systèmes automatisés de détection de contenus illicites. Le champ d'application de ces outils s'est élargi grâce à l'émergence de modèles d'apprentissage automatique *(machine learning)*, capables d'identifier des images et textes plus complexes, de comprendre le contexte d'une phrase ou d'une image, voire de juger de la véracité d'une affirmation.

Le futur règlement européen Digital Services Act (DSA) met en lumière les multiples rôles d'algorithmes dans la gestion de contenus sur les réseaux sociaux. Ces algorithmes identifient des contenus illicites et procèdent à leur retrait avec ou sans intervention humaine ; ils signalent l'existence d'utilisateurs potentiellement abusifs du service ; ils organisent la présentation de contenus et de publicités aux utilisateurs en fonction de leurs profils. Le règlement DSA propose d'encadrer l'utilisation d'algorithmes, surtout ceux utilisés pour retirer des contenus illicites. Les outils sont calibrés pour bloquer automatiquement, et sans intervention humaine, des contenus les plus manifestement illégaux. En cas de doute, la machine enverra le cas à des décisionnaires humains. Une grande partie des décisions de retrait de contenus sont aujourd'hui prises sans intervention humaine (3), ce qui soulève la question de leur légitimité et des garanties qui les entourent. Le DSA prévoit des garanties procédurales et de transparence similaires à celles qui existent pour les décisions prises par l'Etat. Le droit constitutionnel impose à l'Etat des règles contraignantes en matière de blocage de contenus illicites, alors que les plateformes, elles, ne sont pas directement concernées par ces contraintes constitutionnelles. Cependant, les plateformes dites « *structurantes* » ont un pouvoir quasi-étatique en matière de liberté d'expression. Il est donc logique d'étendre à ces plateformes les règles de transparence et de procédure qui s'appliquent aux décisions de l'Etat.

En 2018, les organisations de défense des droits civiques aux Etats-Unis ont élaboré des principes minimaux de transparence et de procédure équitable qui doivent s'appliquer aux décisions de retrait de contenus ou de suspension de comptes sur les réseaux sociaux. Appelés « Santa Clara Principles » (4), ces principes non-contraignants recommandent la publication par chaque plateforme numérique de données détaillées sur les alertes, les décisions de retrait et de suspension. Ils prévoient la notification aux utilisateurs affectés par les décisions de retrait, la publication de règles claires sur les types de contenus interdits sur la plateforme, la mention de raisons du retrait, la fourniture d'informations sur l'utilisation ou non d'un outil automatique, et une procédure efficace de contestation devant un décisionnaire humain différent de la personne qui a pris la décision initiale. Les Santa Clara Principles (SCP) reprennent, pour les adapter aux plateformes, une partie des règles constitutionnelles de « *due process* » aux Etats-Unis qui s'appliquent aux décisions, notamment algorithmiques, de l'Etat.

Le DSA va plus loin que les « SCP »

Le projet de règlement DSA rendrait contraignant un certain nombre des SCP, et notamment l'obligation d'informer l'utilisateur que son contenu a été retiré et de lui fournir une explication sur les raisons du retrait. La notification doit également mentionner l'utilisation éventuelle d'un outil automatique, et fournir des informations claires sur la possibilité de contester la décision. Le DSA exige une procédure efficace pour gérer les contestations d'utilisateurs, une procédure qui ne peut pas s'appuyer uniquement sur des moyens automatisés. Les utilisateurs peuvent donc contester un retrait devant un décisionnaire humain. Le DSA

Notes

(1) - Soumettre la publication de contenus à un contrôle préalable est généralement considéré comme une atteinte disproportionnée à la liberté d'expression (en France, Nouveaux Cahiers du Conseil constitutionnel n°36, juin 2012 ; aux Etats-Unis, New York Times Co. v. United States, 403 U.S. 713, 1971).

(2) - Directive européenne « E-commerce » (2000/31).

(3) - Cambridge Consultants, Use of AI in Online Content Moderation, Report for OFCOM, 18-07-19, p. 36.

(4) - https://santaclaraprinciples.org

va au-delà des SCP en matière de transparence algorithmique, en exigeant la publication par les plateformes structurantes d'information sur les objectifs poursuivis par l'algorithme, les indices de performance, et les garanties entourant son utilisation.

Le projet de loi français sur le *« respect des principes de la République »*, adopté par l'Assemblée nationale le 16 février dernier et actuellement examiné au Sénat (5), va plus loin encore en prévoyant la communication au Conseil supérieur de l'audiovisuel (CSA) des paramètres utilisés par les outils automatisés, des méthodes et des données utilisées pour l'évaluation et l'amélioration de leur performance.

Algorithmes, *« faux positifs »* et censure

La performance des algorithmes sera un sujet-clé pour le régulateur. Quel est le niveau acceptable de *« faux positifs »*, à savoir des contenus bloqués à tort ? On sait que les tribunaux n'apprécient guère les faux positifs en matière de liberté d'expression *(lire encadré ci-dessous)* et qu'un algorithme d'apprentissage automatique va forcément générer des faux positifs. Le niveau de faux positifs dépendra notamment du niveau de sensibilité de l'algorithme dans la détection de « vrais » positifs, par exemple une vraie vidéo terroriste. Si l'on réduit le nombre de faux positifs, on va nécessairement réduire la sensibilité de l'algorithme dans la détection de vrais cas de contenus illégaux. Le bon équilibre entre les faux positifs et les faux négatifs sera un sujet délicat, et le niveau d'équilibre sera différent selon le type de contenus. Laisser passer la vidéo d'un acte terroriste du type Christchurch aura un coût très élevé pour la société, alors que laisser passer un morceau de musique protégé par le droit d'auteur sera *a priori* moins dommageable.

Les taux d'erreurs algorithmiques peuvent varier en fonction de la langue utilisée – un algorithme d'analyse de textes sera généralement plus performant en anglais – et peuvent également refléter les biais présents dans les données d'entraînement. Les algorithmes apprennent à partir des exemples de contenus retirés précédemment par les analystes humains. Ces analystes humains sont faillibles. Ils ont leur propre biais – biais culturels, linguistiques, ethniques, de genre – et commettent eux-aussi des erreurs d'appréciation qui seront reproduits ensuite par les algorithmes (6). Ainsi, il faut veiller non seulement au « bon » niveau de faux positifs et de faux négatifs selon le type de contenu, mais également vérifier que le niveau de perfor-mances de l'algorithme ne varie pas selon la couleur de la peau ou le sexe des personnes impliquées, selon la langue utilisée, ou selon le type de discours haineux (7). Ces multiples équilibres devraient être abordés dans un premier temps dans les études de risques systémiques conduites par les plateformes structurantes, en application de l'article 26 du futur règlement DSA en Europe. Ces études devront analyser l'impact des algorithmes d'identification et de retrait de contenus sur les droits fondamentaux. Ainsi, les plateformes devront proposer des solutions techniques et humaines pour concilier des objectifs – souvent contradictoires – liés à la mise en place d'un système de détection performant qui respecte en même temps la liberté d'expression, la protection des données personnelles et la protection contre les discriminations. Actuellement, le projet de règlement DSA prévoit que la Commission européenne sera le régulateur principal pour les plateformes structurantes. Celle-ci pourra émettre des recommandations relatives aux systèmes algorithmiques. Mais la manière de gérer les tensions entre la liberté d'expression et d'autres droits est avant tout une affaire nationale, dépendant du contexte, de l'histoire et de la culture de chaque pays (8).

En France, le CSA serait mieux placé que la Commission européenne pour évaluer les systèmes algorithmiques mis en place par les grandes plateformes pour analyser des contenus destinés au public français. Le paramétrage des algorithmes devra nécessairement refléter ces circonstances locales, et le contrôle de ces paramètres relèverait plus naturellement de la compétence du régulateur national. Un contrôle national de ces outils renforcerait en revanche le morcèlement des approches réglementaires entre Etats membres, et nécessiterait donc un système de coordination au niveau européen similaire à ce qui existe pour la régulation des télécoms et le RGPD. @

** Winston Maxwell, ancien avocat, est depuis juin 2019 directeur d'études Droit et Numérique à Telecom Paris.*

Notes

(5) - https://lc.cx/PL-Sénat-RPR

(6) - Binns R., Veale M., Van Kleek M., Shadbolt N. (2017) Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. In: Ciampaglia G., Mashhadi A., Yasseri T. (eds) Social Informatics. SocInfo 2017. Lecture Notes in *Computer Science*, vol 10540. Springer, Cham.

(7) - Gorwa R, Binns R, Katzenbach C. Algorithmic content moderation. Big Data & Society. January 2020.

(8) - CJUE, Affaire C-73/07.

Focus

Le droit est allergique aux surblocages

Le droit constitutionnel est peu tolérant aux *« faux positifs »* en matière de liberté d'expression. Les risques de surblocage ont été soulignés par la Cour suprême des Etats-Unis dans l'affaire « Reno c. ACLU » (1) dans les années 1990, et par la Cour de justice de l'Union européenne (CJUE) dans les affaires « Scarlet c. Sabam » (2) en 2011 et « Sabam c. Netlog » (3) en 2012. Ces deux dernières affaires concernaient la mise en place, à la demande d'un tribunal belge, d'un dispositif simple pour bloquer des contenus protégés par le droit d'auteur, s'appuyant sur un procédé de *« hash »* pour identifier les fichiers contrevenants.

La CJUE a considéré que ce procédé créait une atteinte disproportionnée à la protection des données à caractère personnel, mais également à la liberté d'expression en raison du risque de surblocage. L'outil serait incapable de détecter s'il s'agissait d'une citation, d'une parodie ou d'une autre utilisation permises par l'une des exceptions du droit d'auteur.

Plus récemment, le Conseil constitutionnel a annulé deux dispositions de la loi française « Avia » (contre la cyberhaine) en raison du risque de surblocage de contenus *« non manifestement illicites »* (4). Pour des contenus faisant l'apologie du terrorisme, le Conseil constitutionnel a considéré que les injonctions de l'autorité administrative (5) ne constituaient pas une garantie suffisante et que les opérateurs de plateformes ne devaient pas suivre ces injonctions de manière automatique. @

1 - Affaire « Reno v. American Civil Liberties Union », *521 U. S. 844 (1997) p. 874 • 2 - Affaire C-70/10, 24-11-11, point 52 • 3 - Affaire C-360/10 du 16-02-12, point 50. • 4 - Décision n° 2020-801 DC du 18 -06-20, point 19. • 5 - Commentaire de la décision n° 2020-801 DC du 18-06-20, page 16.*