Les algorithmes des plateformes

I. Des algorithmes souvent opaques et dont la neutralité est questionnable

1. Définitions : modération versus recommandation

Chaque jour, plusieurs dizaines voire centaines de millions de contenus (textes, vidéos, photos…) sont publiés sur les réseaux sociaux. Pour ordonner cette masse gigantesque, les opérateurs de plateformes numériques ont recours à des algorithmes. Ceux-ci sont principalement de deux types.

On trouve d'une part les **algorithmes de modération** qui ont pour objectif de **repérer les contenus manifestement illicites ou contrevenant aux conditions générales d'utilisation** de la plateforme et qui seront donc supprimés, démonétisés ou rendus moins visibles (on parle d'obfuscation ou de *shadow-banning*). **Cela concerne les publications mais aussi les commentaires des utilisateurs**. Cette modération peut avoir lieu ***ex ante***, c'est-à-dire avant même la publication, pour les contenus les plus graves en particulier, comme la pédopornographie[1]. **Cette détection peut être entièrement algorithmique ou peut consister en une pré-détection qui est ensuite raffinée par des modérateurs humains** (dont les conditions de travail ont été beaucoup décrites et décriées ces dernières années[2]).
D'après le bilan 2021 de l'Arcom sur les moyens et mesures mis en œuvre par les opérateurs de plateforme en ligne pour lutter contre la manipulation de l'information, *"si la majorité des opérateurs (Meta, Pinterest, Twitter, Snap, TikTok) font mention dans leurs déclarations du recours à des outils automatisés pour identifier ou traiter des fausses informations, seul Meta apporte des précisions sur le fonctionnement des outils de détection*."[3] Toutefois, Meta ne fait pas état de l'efficacité de ces outils d'étiquetage des fausses informations, de reconnaissance des contenus déjà modérés et de détection automatique des variations de ce contenu et des risques de faux positifs et donc de "sur-modération" de ces derniers. L'Arcom précise que "*si plusieurs opérateurs, comme Twitter, justifient ce choix par la volonté de ne pas donner de détails pour des raisons de sécurité et empêcher un détournement de leurs services, il subsiste un manque flagrant de transparence en la matière. Néanmoins, cela peut notamment s'expliquer par des approches différentes en matière de lutte contre la manipulation de l'information, certains opérateurs indiquant concentrer leur approche sur la détection de comportements ou phénomènes pouvant porter atteinte à l'intégrité des services*."

Il faut rappeler que **la classification des contenus informationnels est particulièrement difficile et implique de garantir un équilibre parfois ténu avec la liberté d'expression**. **Certaines plateformes trouvent ainsi un contournement, avec la modération communautaire** notamment. Celle-ci est au cœur du processus d'évaluation des contenus de la plateforme Wikipédia. La plateforme X propose

---

[1] https://observatoire-ia.pantheonsorbonne.fr/actualite/moderation-haine-en-ligne-et-lintelligence-artificielle

[2] https://www.francetvinfo.fr/replay-radio/en-direct-du-monde/meta-devant-la-justice-kenyane-des-moderateurs-de-contenus-de-la-maison-mere-de-facebook-denoncent-des-conditions-de-travail-indignes_5795960.html ; https://www.radiofrance.fr/franceculture/daniel-motaung-il-faut-professionnaliser-le-metier-de-moderateur-de-contenus-2237136 ; https://www.la-croix.com/Economie/On-voyait-tout-temps-gens-morts-moderateur-reseaux-sociaux-eprouvant-travail-lombre-2022-01-30-1201197579

[3] https://www.arcom.fr/nos-ressources/etudes-et-donnees/mediatheque/lutte-contre-la-manipulation-de-linformation-sur-les-plateformes-en-ligne-bilan-2021

aussi aux utilisateurs de rédiger des “Notes communautaires” (*community notes*). Pour ce faire, les utilisateurs doivent s'inscrire en tant que contributeurs. Par la suite, lorsqu'un post comporte une information erronée, un contributeur peut proposer une note précisant le contexte, le caractère trompeur ou erroné, ajoutant une source… Cette note est ensuite revue par les autres contributeurs : si elle réunit suffisamment d'avis positifs de leur part, estimant qu'elle est utile à une meilleure compréhension du post, la note est publiée en encadré sous le post, visible pas tous.

Si la modération par la suppression des posts est souvent critiquée, **elle peut également passer par la réduction de la visbilité de ceux-ci, stratégie dont les médias se plaignent particulièrement** depuis l'entrée en vigueur des droits voisins. En d'autres termes, les plateformes exercent un réel rôle de contrôleurs d'accès (*gatekeepers*) en configurant qui est exposé à quelle information. Pour Romain Badouard, “***c'est quelque chose de nouveau dans l'histoire : on ne décide pas de qui émet mais de qui est exposé à l'information***.”[4]

D'autre part, les opérateurs de plateformes numériques ont recours à des **algorithmes de recommandation qui visent à amplifier la visibilité d'un contenu**. Si le débat public s'est beaucoup concentré sur les algorithmes de modération, les algorithmes de recommandation jouent un rôle particulièrement crucial. **Ils déterminent quels contenus émergeront de la masse des millions de publications quotidiennes et ce qui restera dans l'ombre**. Ces algorithmes sont plus ou moins récents selon les plateformes : Facebook l'instaure dès 2006 tandis qu'il n'arrivent qu'en 2016 sur Twitter, devenu X[5]. Ces algorithmes se nourrissent à la fois des traces que nous laissons (likes, commentaires, partages, vidéos lues…) et de celles de nos proches (abonnés, utilisateurs aux caractéristiques similaires, utilisateurs d'un même pays…). Ils sont également en constante évolution en fonction des performances de la plateforme (voir la fiche dédiée aux modèles économiques des plateformes). **En conséquence, le contenu présenté est différent pour chaque utilisateur, il n'y a pas deux fil de contenus identiques**. À cela s'ajoute le fait que **toutes les plateformes n'ont pas les mêmes critères dans leurs algorithmes** : Facebook priorise historiquement les contenus partagés ou appréciés par les proches d'un utilisateur, dans la continuité de son offre de service qui était de garder le contact avec ses amis et sa famille, tandis que Twitter - devenu X - mise historiquement davantage sur les contenus politiques[6].

2. Les algorithmes sont-ils la cause de bulles de filtre et de chambres d'échos ?

Parce que les algorithmes de recommandations se nourrissent de nos interactions passées sur la plateforme et de celles des comptes qui nous suivent et que nous suivons, certains travaux alertent sur le risque que se forme une boucle auto-alimentée finissant par suggérer sans cesse le même type de contenus. On parle de **bulle de filtre**[7] ou **chambre d'écho**. Toutefois, **cette notion fait l'objet de débat**, certains arguant que la bulle de filtre est une construction algorithmique, tandis que d'autres mettent en lumière la diversité des informations présentées en ligne et le caractère davantage choisi des bulles de filtre[8].

Dans son ouvrage *Toxic Data*, David Chavalarias rappelle l'intensité des chambres d'échos dans la période de l'épidémie de Covid-19 : “*Dans telle chambre d'écho, on croyait dur comme fer aux bienfaits de l'hydroxychloroquine comme traitement préventif ; dans celle d'en face, on ne jurait que par les*

[4] Audition de Romain Badouard par les EGI.
[5] Toxic Data
[6] Toxic Data
[7] Eli Pariser.
[8] Ce choix est notamment dicté par le biais de confirmation.

vaccins. [...] Ces chambres d'écho numériques ont favorisé la survenue de situations extrêmes à travers le monde. Des sous-populations ont développé des croyances déconnectées des faits admis par les scientifiques et par le reste de la société (souvenez-vous : le masque est inefficace et dangereux pour la santé, les vaccins contiennent des puces 5G, etc.)."[9] Pour lui, "*les chambres d'échos se forment plus facilement en ligne que hors ligne - un phénomène qui, hélas, réduit la circulation d'informations au sein de la société, et contribue à fragmenter les groupes sociaux numériques. La façon dont nous faisons société en ligne est donc très différente de celle dont nous faisions société avant l'apparition du numérique*."[10]

Il est à noter que **cet enjeu des bulles de filtres ne se présente pas uniquement pour les algorithmes de recommandation, mais également les résultats de recherche**. Les analyses de Tracking.Exposed ont montré que les résultats de recherche sont également personnalisés en fonction de l'historique de l'utilisateur. En investiguant la plateforme YouTube pendant l'élection présidentielle américaine de 2020, le collectif a observé que les requêtes "*election results*" n'obtenaient pas les mêmes résultats en fonction de l'orientation politique des utilisateurs. Ceux proches des Démocrates obtenaient des résultats acclamant la victoire de Joe Biden, tandis que ceux proches de Républicains avaient davantage de contenus dénonçant la fraude électorale : "*C'est une bulle de filtre particulièrement insidieuse car, si les utilisateurs peuvent s'attendre à ce que leurs recommandations soient personnalisées, ils s'attendent à ce que leurs résultats de recherche soient plus universels. Le résultat est que chacun vit dans l'illusion que sa propre opinion est la vision dominante et légitime.* »[11]

Ces effets de bord des algorithmes de recommandation ne sont pas neutres. Au lendemain de l'élection de Donald Trump en 2016, une enquête menée par le Pew Research Center à montré que parmi les personnes ayant eu des discussions politiques sur les réseaux sociaux, 59 % les ont trouvées "stressantes et frustrantes" et 64 % en sont ressorti avec le sentiment d'avoir moins de choses en commun avec leur interlocuteur qu'elles ne le pensaient[12]. Cette tension croissante en ligne pousserait même les utilisateurs à fuir les contenus qu'ils trouvent offensants ou controversés, comme cela a été rapporté par 83 % des enquêtés. Une autre étude américaine a montré que l'usage quotidien de Facebook et Twitter est associé à la perception d'un désaccord plus important avec les personnes que nous fréquentons dans le monde physique[13]. Et David Chavalarias de conclure "***Nos interactions virtuelles polluent donc nos interactions dans le monde réel !***"[14].

Toutefois, **d'autres enquêtes nuancent ces résultats**. Barberá et al. (2015) montrent ainsi que si pour les événements politiques les échanges sur les réseaux sociaux ont principalement lieu entre personnes partageant les mêmes préférences idéologiques, il en va autrement des autres événements d'actualité, comme les attentats de Boston en 2013 par exemple, ou le Super Bowl[15]. Dans ce second cas, ils observent un processus dynamique partant d'interactions inter-idéologiques avant de se transformer en débat polarisé. Pour les auteurs, le degré de ségrégation idéologique sur les réseaux sociaux aurait donc été surestimé. **Ils qualifient ainsi les échanges sur ces plateformes de**

---

[9] Toxic Data, pp. 59-60.
[10] *Ibid.*, p. 60-61.
[11] Marc Faddoul cité par https://cnnumerique.fr/files/uploads/2022/Livres/CNNum-recits-et-contre-recit-itineraire-des-fausses-informations-en-ligne.pdf
[12] https://www.pewresearch.org/internet/2016/10/25/the-political-environment-on-social-media/ cité dans Toxic Data p. 61.
[13] HAMPTON, Keith N., SHIN, Inyoung, et LU, Weixu. Social media and political discussion: when online presence silences offline conversation. *Information, Communication & Society*, 2017, vol. 20, no 7, p. 1090-1107, cité par Toxic Data p. 61.
[14] *Toxic Data* p. 63.
[15] Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber?. *Psychological science, 26*(10), 1531-1542

“conversation nationale”. De la même façon, Dutton et al. (2017) insistent sur le fait que, dans les sept pays de leur étude (Grande-Bretagne, France, Allemagne, Italie, Pologne, Espagne et États-Unis), 36 % des utilisateurs d’Internet sondés sont exposés « souvent ou très souvent » à une grande variété de points de vue et d’opinions et 43 % « de temps en temps »[16].

En France, les travaux actuellement menés en économie par Marianne Lumeau, Stéphanie Peltier, Sylvain Dejean et Benoît Tarroux[17] convergent vers ces résultats. Ces chercheurs ont proposé à un échantillon d’individus différents menus de médias, soit avec uniquement des médias de gauche, soit uniquement de droite, soit les deux et en faisant varier l'extrémisme des médias (par exemple en ajoutant Valeurs Actuelles à certains menus), le nombre de médias dans chaque menu et le contexte du choix (est-ce que l’individu choisit pour lui-même ou pour toute la société). Il apparaît que lorsque l’individu décide pour lui-même, 5 groupes se forment[18] :

Types de clusters		Fréquence (en %)
Préférence pour soi-même (*Me*)		
Préférence pour la bulle d’information	Menus de médias alignés sur leur propre opinion politique	42,2
Préférence pour le choix	Menus de médias diversifiés quand ajouts d’une option mais alignés sur leur opinion politique à taille constante	20,6
Préférence pour la diversité avec aversion pour les extrêmes	Menus de médias diversifiés sans opinions politiques extrêmes	17,7
Préférence pour une bulle élargie partisane	Menus de médias étendus à l’extrême du même bord politique	14,0
Indifférent	Aucune préférence pour un menu de médias particulier	5,4

Méthode : Classification par la méthode k-modes avec k = 5
Lecture : 42,2% des individus ont une préférence pour la bulle lorsqu’ils choisissent un menu de médias pour eux-mêmes

Il est intéressant de noter que ces choix varient lorsque l’individu doit choisir non plus pour lui mais pour l’ensemble de la société. La répartition se fait alors de la sorte :

Types de clusters		Fréquence (en %)
Préférence pour la société (*Library*)		
Préférence pour la bulle d’information	Menus de médias alignés sur leur propre opinion politique	35,0
Préférence pour la diversité avec aversion pour les extrêmes	Menus de médias diversifiés sans opinions politiques extrêmes	26,0
Préférence pour le choix	Menus de médias diversifiés quand ajouts d’une option mais alignés sur leur opinion politique à taille constante	21,8
Préférence pour la diversité	Menus de médias diversifiés politiquement y compris avec les opinions extrêmes	12,0
Indifférent	Aucune préférence pour un menu particulier	5,2

Méthode : Classification par la méthode k-modes avec k = 5
Lecture : 35% des individus ont une préférence pour la bulle lorsqu’ils choisissent un menu de média pour les autres.

[16] Dutton, W. H., Reisdorf, B., Dubois, E., & Blank, G. (2017). Search and politics: The uses and impacts of search in Britain, France, Germany, Italy, Poland, Spain, and the United States.
[17] Intervention ARCOM Research Day (16/11/2023) - A preference for filter bubbles?
[18] Résultats présentés à l’occasion de la Journée d’études 2023 de l’Arcom du 16 novembre 2023.

Quels sont les déterminants des trois clusters dominants ? Leurs résultats montrent que **la probabilité d'avoir une préférence pour la bulle de filtre est significativement plus élevée quand on choisit pour soi-même plutôt que pour la société**. **Ceux qui préfèrent la bulle de filtre sont plutôt des personnes de plus de 65 ans, généralement se déclarant de droite et légèrement plus souvent des femmes**. En revanche, ceux qui sont plus enclins à privilégier le choix sont plutôt des personnes de 18 à 34 ans, plutôt des hommes et généralement se déclarant de gauche. Le niveau d'éducation ne joue un rôle significatif que pour la troisième catégorie : ce sont les personnes ayant au moins le bac qui sont les plus susceptibles d'avoir une appétence pour la diversité, sauf pour les extrêmes. Cette attitude est aussi davantage observée lorsqu'il s'agit d'un choix pour la société et non pour soi-même et pour les personnes se déclarant comme centristes.
Ces résultats sont corroborés par les analyses empiriques menées par AI Foresics qui montrent que les utilisateurs s'abonnent davantage à des comptes proches de leur sensibilité politique et interagissent davantage avec des posts congruents avec leur positionnement politique.

Au-delà de la recommandation de contenus pour coller au mieux à ce que la plateforme suppose que les utilisateurs viennent chercher sur leur service, beaucoup de travaux se sont interrogés sur la neutralité de ces algorithmes. Pour le dire autrement, **ces algorithmes favorisent-ils certains contenus plutôt que d'autres et si oui, lesquels et pourquoi ?**

3. Les algorithmes de recommandations sont-ils neutres ?

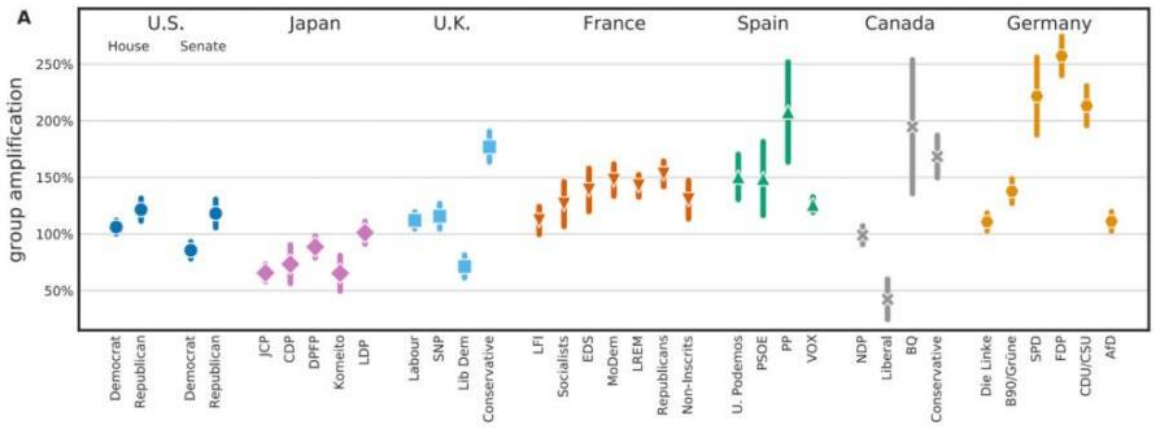
En premier lieu, importe de se demander **dans quelle mesure ces algorithmes de recommandation impactent-ils ce à quoi les utilisateurs sont exposés ?** Le projet CrossOver réunit l'EU Disinfo Lab, Apache, SavoirDevenir et CheckFirst. Il vise à suivre et examiner l'influence des algorithmes de recommandations sur les principaux réseaux sociaux, notamment sur la propagation de fausses informations. Pour ce faire, le projet simule des utilisateurs neutres et observe ensuite les contenus recommandés et leur évolution dans le temps. Ces données sont comparées à celles fournies par les API des plateformes - pour celles qui le proposent. Actuellement, le projet se concentre sur la Belgique, mais est en train de s'étendre à sept nouvelles régions francophones : la France, le Mali, le Maroc, le Québec, le Sénégal, la Suisse et la RDC. **Ce travail montre le rôle important des algorithmes de recommandation et de référencement dans l'information relayée aux utilisateurs, y compris les résultats des requêtes et l'autocomplétion des barres de recherche**. Par exemple, lorsque le mot "Donbass" est tapé dans la barre de recherche Google, l'algorithme de prédiction propose "Donbass Insider", un média en ligne francophone pro-Kremlin créé en 2016[19]. Sur Google News, une étude du collectif a montré que les recherches réalisées en Belgique sur la province du Xinjiang conduisait en partie (13 % des articles proposés) à des contenus produits par six acteurs contrôlés par le gouvernement de la RPC, propageant un narratif extrêmement positifs autour de cette province et ne faisant pas ou très peu mention de la communauté Ouïghour et des enjeux de droits humains qui s'y présentent[20]. Un résultat très similaire a été observé sur YouTube : après le bannissement de Russia Today de la plateforme, les contenus proposés lors de requêtes sur la Russie ont été progressivement remplacés par des vidéos de chaînes contrôlées par le gouvernement chinois[21].

Sans être des algorithmes de recommandation à proprement parler, les algorithmes peuvent être utilisés comme outils de régulation des contenus publicitaires en ligne avec des effets de bord là aussi. La publicité peut en effet être instrumentalisée par les acteurs propageant de fausses informations, on parle de "***dark ads***". À titre d'exemple, lors de la campagne précédant le vote sur le Brexit, la campagne *VoteLeave* (pro-Brexit) avait financé des publicités sur Facebook propageant des informations erronées sur le coût hebdomadaire de l'adhésion à l'Union européen,nes ou encore sur

[19] https://crossover.social/disinformation-on-donbas-is-only-a-google-autocomplete-away/
[20] https://crossover.social/is-china-trying-to-control-the-narrative-about-xinjiang-on-google-news/
[21] https://crossover.social/are-youtube-algorithms-addicted-to-state-controlled-media/

l'entrée prochaine de la Turquie et de l'Albanie dans l'Union[22]. **Pour les détecter, des algorithmes sont utilisés, mais ils ne sont pas immunes d'erreur.** Grazia Cecere et Clara Jean ont ainsi montré que ces algorithmes peuvent entraîner de mauvaises qualifications des contenus publicitaires. Certaines publicités sont donc par erreur disqualifiées et ne sont plus affichées aux utilisateurs. Paradoxalement, ces erreurs de classification se constatent proportionnellement plus fréquemment lorsqu'il s'agit d'organisations institutionnelles qui sont pourtant moins susceptibles de propager des fausses informations[23].

Il semble donc que **les contenus recommandés par les plateformes ne se basent pas uniquement sur les appétences de l'utilisateur, ses recherches antérieures et celles de ses proches, mais que d'autres facteurs entrent en ligne de compte. On peut donc se demander si cette recommandation est biaisée et, si oui, en faveur de quels groupes ?**
Dans une perspective de rétro-ingénierie, Twitter a publié en octobre 2021 une étude sur l'amplification algorithmique des messages publiés sur le réseau social par des personnalités politiques dans sept États (l'Allemagne, le Canada, l'Espagne, les États-Unis, la France et le Royaume-Uni). Il ressort assez clairement de leur analyse que l**es publications des personnalités de partis de droite sont plus amplifiées que celles provenant de personnalités de gauche, à l'exception de l'Allemagne.** Il en va de même pour les médias[24]. En France, c'est donc la parole des Républicains qui est particulièrement diffusées avec une amplification 2,5 fois plus importante que des contenus neutres.



Source : HUSZÁR, Ferenc, KTENA, Sofia Ira, O'BRIEN, Conor, *et al.* Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences*, 2022, vol. 119, no 1, p. e2025334119.

Cependant, **Twitter reconnaît ne pas connaître la cause de cet écart** : "*Il est beaucoup plus difficile de déterminer pourquoi ces modèles observés se produisent, car il s'agit d'un produit des interactions entre les personnes et la plateforme*"[25]. La plateforme invite ainsi à poursuivre l'analyse des racines de ce phénomène afin de déterminer s'il est nécessaire de modifier l'algorithme afin de réduire ses effets négatifs.

Quels pourraient être les facteurs explicatifs de cette situation ? David Chavalarias propose trois éléments : des différences de tailles de populations en fonction des bords politiques ; des différences de sensibilité moyenne au biais de confirmation et au biais de négativité en fonction des préférences

[22] https://www.joe.co.uk/politics/brexit-facebook-adverts-192164
[23] CECERE, Grazia, JEAN, Clara, LEFRERE, Vincent, *et al. Tradeoffs in automated political advertising regulation: evidence from the COVID-19 pandemic*. 2020.
[24] Turcan, M. "Twitter admet qu'il amplifie plus la droite française que la gauche". *Numerama*. 22 octobre 2021.
[25] HUSZÁR, Ferenc, KTENA, Sofia Ira, O'BRIEN, Conor, *et al.* Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences*, 2022, vol. 119, no 1, p. e2025334119.

politiques ; une différence de rhétorique entre les élus de gauche et de droite et notamment leur façon de recourir à des messages négatifs particulièrement viraux au regard du biais de négativité (voir fiche sur le modèle économique des plateformes)[26]. Jen Schradie éclaire elle aussi cette asymétrie entre partis progressistes et conservateurs. Dans son ouvrage *L'Illusion de la démocratie numérique : Internet est-il de droite ?* (2022), elle retrace son étude sur quatre ans d'une quarantaine de groupes politiques en Caroline du Nord, allant de l'extrême droite à l'extrême gauche pour mieux comprendre leurs usages d'Internet. Elle en conclut que les idées conservatrices sont aujourd'hui dominantes sur Internet. Pour elle, plusieurs éléments cumulatifs peuvent expliquer ce constat :

1. **Les fractures sociales se reproduisent sur Internet** : les contenus les plus viraux en ligne sont des contenus "fabriqués" (*manufactured*), c'est-à-dire soit produits par des personnes rémunérées ou bénévoles qui travaillent à temps plein sur cette tâche et qui tentent de comprendre le fonctionnement des algorithmes des plateformes et leurs évolutions pour avoir une meilleure visibilité en ligne. Ces producteurs sont en majorité les personnes qui sont également dominantes dans la vie physique (les classes sociales éduquées, riches, blanches…) ;
2. **Pour avoir un impact en ligne, les collectifs doivent être organisés et hiérarchisés**, ce qui est traditionnellement plutôt le cas des mouvements conservateurs, les mouvements progressistes se caractérisant davantage par une horizontalité et souvent en éphémérité. La parole est diluée entre un plus grand nombre de personnes et les membres des collectifs progressistes sont souvent moins actifs en ligne ;
3. **Les contenus viraux sur Internet sont les contenus polémiques et synthétiques**. Là encore, les thématiques promues par les conservateurs sont souvent très précises et polémiques (comme les médias, l'immigration…). En revanche, à gauche, les thèmes sont plus larges et déclinés sur davantage de sujets différents (par exemple le thème de la justice concerne à la fois les droits des LGBT +, le féminisme, la justice sociale…). Il est donc plus difficile de résumer une pensée unifiée en 280 caractères.

En France, Tim Faverjon et Pedro Ramaciotti-Morales ont montré que les algorithmes de recommandation des plateformes sont fortement corrélées avec les attitudes politiques des utilisateurs. Cette observation est d'autant plus forte que les utilisateurs sont situés à des extrêmes politiques. En d'autres termes, **ces algorithmes sont capables d'apprendre indirectement les opinions politiques des utilisateurs et de les utiliser pour lui faire des recommandations, sans que cela ne soit explicitement demandé**[27].

En ce qui concerne l'invisibilisation de certains contenus, Romain Badouard souligne que ces stratégies ne sont pas utilisées pour promouvoir des contenus de qualité, mais **à des fins politiques**[28]. Par exemple, Tik Tok est accusé d'instrumentaliser le mouvement Black Lives Matter aux États-Uni[29]s : dès que le #BLM est apposé, le contenu n'est plus mis en avant sur la page "For you". Au-delà de cet exemple, c'est un problème plus global qui se pose sur ce réseau social : la plateforme est accusée d'invisibiliser les mouvements défendant les minorités en général[30]. Les réseaux sociaux du groupe Meta ont aussi été visés par plusieurs accusations à ce sujet comme les activistes féministes à propos

---

[26] *Tixuc Data*, p. 91.

[27] https://medialab.sciencespo.fr/activites/ai-political-machine/. Travaux présentés à l'occasion de la journée d'études de l'Arcom du 16 novembre 2023.

[28] Audition de Romain Badouard par les EGI.

[29] https://time.com/5863350/tiktok-black-creators/

[30] https://www.theguardian.com/technology/2020/mar/17/tiktok-tried-to-filter-out-videos-from-ugly-poor-or-disabled-users ; https://www.buzzfeednews.com/article/laurenstrapagiel/tiktok-algorithim-racial-bias

d'Instagram ou les Gilets Jaunes se plaignaient à propos de Facebook[31]. Malgré cela, les textes européens (voir partie III) ne comportent que peu de choses au sujet du *shadow-banning*

Au-delà des sympathies politiques, **il demeure plus généralement une très grande opacité quant aux critères retenus dans le fonctionnement des algorithmes de recommandation et pourquoi**. Ce manque de transparence est d'ailleurs régulièrement pointé par l'Arcom dans ses bilans annuels des moyens et mesures mis en œuvre par les opérateurs de plateforme en ligne en France pour lutter contre la manipulation de l'information[32]. Dès 2019, dans sa recommandation n°2019-03 du 15 mai 2019, l'Arcom (encore appelé CSA) appelait les plateformes "*à fournir une information claire, suffisamment précise et facilement accessible sur les critères ayant conduit à l'ordonnancement du contenu proposé à l'utilisateur et le classement de ces critères selon leur poids dans l'algorithme.*"[33] **Cet effort de transparence est également un engagement pris par les signataires du Code européens de bonnes pratiques renforcé contre la désinformation**[34].
Toutefois, certaines plateformes ont transmis à l'Arcom des éléments éclairant au moins partiellement leurs algorithmes de recommandation :

Extrait du Bilan 2021 des moyens et mesures mis en œuvre par les opérateurs de plateforme en ligne pour lutter contre la manipulation de l'information publié par l'Arcom le 28 novembre 2022

Afin d'effectuer la tâche ou série de tâches pour laquelle ils ont été programmés, les algorithmes de recommandation utilisent des données en entrée et produisent un résultat en sortie qui permet la recommandation de contenus. Ainsi, le choix des critères principaux utilisés par les opérateurs et leur équilibrage peuvent être des facteurs expliquant une plus ou moins grande exposition à des fausses informations. À l'exception de LinkedIn et de Pinterest, les opérateurs concernés font état de nombreux éléments sur lesquels ils basent les recommandations.

Concernant les résultats de recherche de contenus, Google indique afficher les meilleurs résultats d'une recherche sur Google Search en fonction des mots-clés de la requête, puis de la recherche de correspondance ainsi que du classement des pages selon leur utilité. Microsoft génère les résultats de recherche sur Bing à l'aide de systèmes de classement relatifs aux mots sur la page ou au titre de la page mise en avant. En outre, les deux opérateurs indiquent se baser sur des critères de pertinence et de fiabilité de l'information pour empêcher la mise en avant de contenus relatifs à de la manipulation de l'information, mais sans préciser comment ils évaluent cette fiabilité.

Le risque de mise en avant de résultats de recherche indexant des contenus de désinformation peut aussi découler d'un phénomène appelé « vides de données » (data voids) qui correspond à l'absence d'informations fiables sur un sujet recherché. Pour y répondre, Google affiche un message informant l'utilisateur qu'il n'a pas été en mesure de trouver de sources fiables correspondantes. Microsoft a lancé des projets de recherche en la matière.

Les résultats de recherche sur YouTube sont basés sur trois notions : pertinence, engagement et qualité. Google indique s'appuyer sur des évaluateurs externes ne précise malheureusement pas les critères d'évaluation lui permettant de juger de la qualité d'un contenu.

Concernant la recommandation sur le fil d'actualité ou la page d'accueil, Meta déclare que la recommandation s'effectue en trois temps : un processus d'intégrité pour réduire le nombre de publications à recommander (sans donner davantage de précisions), la détermination d'un score de qualité et un processus de diversification du type de contenu (contenu écrit, vidéo, issu d'un groupe public ou privé, etc.). Le score de qualité est attribué à chaque contenu sur la base de « milliers de critères » (tels que la mise en avant de contenus non lus, la relation avec le compte partageant le contenu ou encore le type de publication). Meta prend également en compte les types de « réactions » de l'utilisateur face aux contenus pour en pondérer la présentation.

Les contenus recommandés sur Twitter sont majoritairement issus de comptes suivis par l'utilisateur, mais

[31] https://esprit.presse.fr/article/romain-badouard/shadow-ban-l-invisibilisation-des-contenus-en-ligne-43629
[32] https://www.arcom.fr/nos-ressources/etudes-et-donnees/mediatheque/lutte-contre-la-manipulation-de-linformation-sur-les-plateformes-en-ligne-bilan-2021
[33] *Ibid.*
[34] https://digital-strategy.ec.europa.eu/fr/policies/code-practice-disinformation. Engagement 19.

l'opérateur ne précise pas les critères utilisés à l'exception de la popularité ou des interactions avec les autres utilisateurs.

Sur TikTok, la recommandation est en premier lieu basée sur les intérêts exprimés par l'utilisateur lors de l'inscription, sur l'utilisation par ce dernier du bouton « Pas intéressé(e) » et sur les paramètres du compte. La personnalisation du fil se base ensuite sur l'activité de l'utilisateur, telle que les réactions sur les contenus, leur partage, les comptes suivis et les commentaires publiés. Les informations d'une vidéo (hashtags, légendes et son) sont également utilisées pour pondérer l'ensemble de ces données en fonction de la « valeur » d'un contenu, calculée selon la manière qu'ont les utilisateurs de le consommer.

Certains opérateurs ont également fait état de critères utilisés pour des parties spécifiques de leur service. Sur Snapchat, les contenus sont affichés sur la Map en fonction de leur géolocalisation tandis que sur les parties Discover et Spotlight, la recommandation procède des recherches de l'utilisateur, des contenus consommés, des abonnements et des types de lieux visités.

Les recommandations faites à l'utilisateur en train de consommer un contenu sur YouTube (fonctionnalité « À suivre ») se fondent sur plusieurs indicateurs (historique, abonnements, contexte, etc.) mais également sur les usages des autres utilisateurs de la vidéo en cours de visionnage et sur les contenus qu'ils ont consommés par la suite.

L'Arcom s'interroge néanmoins sur les conséquences d'un mauvais équilibrage de l'ensemble ces critères, faute d'informations permettant de comprendre leur hiérarchisation.

L'ensemble de ces éléments concourt à une très grande difficulté de l'audit de ces modèles, notamment pour des tiers indépendants. Par ailleurs, plusieurs des carences précédemment mentionnées sont visées par le règlement européen sur les services numériques (RSN, en anglais *Digital Services Act* - DSA). Cependant, des propositions complémentaires par rapport à ce texte sont également sur la table.

II. Quels leviers face aux biais des algorithmes sur les plateformes numériques ?

1. Que prévoit le RSN en matière d'algorithmes et quelles sont ses limites ?

Le RSN prévoit des dispositions relatives aux algorithmes, mais celles-ci présentent parfois des limites[35] :

- L'**article 24** impose aux fournisseurs de plateformes en ligne de **publier au moins tous les six mois un rapport de transparence** précisant notamment le nombre de notifications reçues par type de contenus présumés illicites, le nombres de notifications reçues de la part de signaleurs de confiance, des informations utiles et compréhensibles sur les activités de modérations, les mesures prises pour dispenser une formation et une assistance aux personnes chargées de la modération des contenus, le nombre et le type de mesures prises qui affectent la disponibilité, la visibilité et l'accessibilité des informations fournies… Faisant suite à cette obligation, X publié le 5 novembre son premier rapport de transparence faisant état de 52 modérateurs francophones.

- L'**article 27** impose aux fournisseurs de plateformes en ligne ayant recours à des systèmes de recommandation d'**établir dans leurs conditions générales**, dans un langage simple et compréhensible, **les principaux paramètres utilisés dans leurs systèmes de recommandation**, ainsi que les options dont disposent les destinataires du service pour modifier ou influencer ces principaux paramètres. Le groupe Meta a d'ores et déjà mis en place

[35] https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX%3A32022R2065

des mesures à cet égard comme le souligne l'Arcom dans son bilan 2021 des moyens et mesures mis en œuvre par les opérateurs de plateforme en ligne pour lutter contre la manipulation de l'information :

"*Seul Meta a mis en place une fonctionnalité contextuelle pour comprendre la recommandation de contenus organiques, appelée « Pourquoi est-ce que je vois ça ? ». Proposé à proximité immédiate du contenu avec un libellé clair, cet outil apporte aux utilisateurs des informations les aidant à comprendre pourquoi un contenu en particulier apparaît sur leur fil, et propose un second niveau d'information en renvoyant vers des éléments plus détaillés sur la recommandation de contenus. Ainsi, les utilisateurs peuvent connaître la fréquence à laquelle ils interagissent avec des publications provenant d'un compte, la popularité de ce dernier et la fréquence à laquelle ils interagissent avec un type spécifique de publication (vidéos, photos ou liens, par exemple). Néanmoins, l'Arcom note que ces informations ne sont pas disponibles pour les contenus relevant de « suggestions » de la plateforme (alors même qu'elles seraient particulièrement utiles) : sur Facebook et sur Instagram, il est uniquement indiqué que les suggestions s'expliquent « en fonction de plusieurs facteurs, dont votre activité sur Instagram » (sans la fourniture d'informations personnalisées supplémentaires).*"[36]

Certaines plateformes proposent également de signaler facilement qu'un contenu ne plaît pas ou que l'on souhaite moins voir de contenus de ce type, comme Instagram, Facebook, TikTok, Snapchat ou Twitter (X).

Toutefois, **cette transparence peut demeurer parcellaire**. Les travaux de Tim Faverjon et Pedro Ramaciotti-Morales montrent que même si de façon explicite les algorithmes ne fonctionnent pas sur la base de données liées à la politique, ils infèrent cette information dans leurs recommandations.

- L'**article 34** prévoit l'**évaluation des risques systémiques** par les fournisseurs de très grandes plateformes en ligne et de très grands moteurs de recherche en ligne liés à la conception ou au fonctionnement de leurs services et de leurs systèmes connexes, **y compris des systèmes algorithmiques**, ou de l'utilisation faite de leurs services.
  De façon complémentaire, **l'article 35 prévoit que ces acteurs mettent en place des mesures d'atténuation raisonnables, proportionnées et efficaces**, adaptées aux risques systémiques spécifiques recensés, incluant des tests et des adaptations de leurs systèmes algorithmiques, dont leurs systèmes de recommandation.

- L'**article 37** impose un **audit indépendant annuel** des fournisseurs de très grandes plateformes en ligne et de très grands moteurs de recherche en ligne, à leur propre frais. **Toutefois, la charge de la sélection de l'auditeur leur incombe et les rapports d'audits ne seront pas nécessairement publics**, limitant la portée de cette obligation.

  Dans son bilan 2021, l'Arcom faisait état de plusieurs dispositifs mis en place par les plateformes pour évaluer en interne ou par un acteur externe leurs systèmes algorithmiques :

  "*Twitter effectue des tests sur la précision de détection d'un outil avant son lancement. Pour ce faire, il l'évalue sur des corpus de contenus issus d'une base de référence en calculant la différence entre le nombre d'actions de modération attendues et le nombre d'actions proposées*

[36] https://www.arcom.fr/nos-ressources/etudes-et-donnees/mediatheque/lutte-contre-la-manipulation-de-linformation-sur-les-plateformes-en-ligne-bilan-2021

par l'outil. Après l'éventuelle modification et le lancement de ce même outil, une équipe dédiée interne à l'opérateur évalue la performance et effectue des diagnostics récurrents.

LinkedIn évalue ses outils en amont de leur mise en production et assure un suivi à la suite de leur déploiement. Sans donner davantage de précisions sur les modalités, Pinterest évoque une amélioration continue de ses modèles par des « analyses hors-ligne » et Snap une régulation humaine de ses outils par un échantillonnage aléatoire. Enfin, Meta déclare développer des outils et processus pour évaluer ses outils et prend pour exemple les progrès réalisés en 2021 en matière de « traitement équitable des communautés » sur ses services (après des audits externes en 2020), exposés dans sa Newsroom."

- L'**article 38** impose aux fournisseurs de très grandes plateformes en ligne et de très grands moteurs de recherche en ligne qui utilisent des systèmes de recommandation de **proposer au moins une option pour chacun de ceux-ci ne reposant pas sur du profilage** (par exemple mais non exclusivement, une présentation chronologique des contenus).

- L'**article 40** dispose que les fournisseurs de très grandes plateformes en ligne et de très grands moteurs de recherche en ligne **donnent accès au coordinateur pour les services numériques de l'État membre d'établissement ou à la Commission**, **à des informations précisant la conception, la logique, le fonctionnement et la procédure de test de leurs systèmes algorithmiques**, y compris leurs systèmes de recommandation. De la même façon, **ces acteurs devront faciliter et fournir l'accès à leurs données à des chercheurs agréés à des fins de recherche** contribuant à la détection, au recensement et à la compréhension des risques systémiques dans l'Union et d'évaluation des mesures d'atténuation des risques.

  Là encore, certaines plateformes ont déjà mis en place un tel dispositif, mais parfois avec quelques limites. Notamment, les données partagées sont souvent parcellaires et ne donnent pas d'information sur la promotion algorithmique et la responsabilité de l'amplification. Il manque également fréquemment de données sur la publicité : qu'est-ce qui est proposé et à qui, sur quelle base ? Quelle vérification y a-t-il de l'information contenue dans les publicités ?

  De son côté, Twitter a longtemps conféré un accès à ses données avant de rendre cet accès payant suite à l'arrivée d'Elon Musk à la tête du réseau social. Ces derniers jours, X a fait évoluer ses conditions d'accès conformément au RSN[37]. Par ailleurs, l'accès à ces API est conditionné au respect de critères et consignes très contraignants pour les chercheurs (informer la plateforme de l'objet de la recherche, obligation de supprimer les données à l'issue etc.).

  Ces contraintes imposent aux chercheurs d'envisager d'autres modes d'audits :

Il existe plusieurs catégories d'audit d'algorithmes :

i) Audit coopératif : les chercheurs recourent à l'API que le DSA impose aux plateformes de créer ;

ii) Audit adversarial (contradictoire) : deux options sont possibles :

- les chercheurs créent de faux comptes qui sont automatisés, mais qui se heurtent cependant aux barrières anti-robots des plateformes ;
- les chercheurs demandent aux utilisateurs de fournir leurs données de consommation, qui donnent une vision plus proche du réel, mais moins généralisable et moins exhaustive.

[37] https://techcrunch.com/2023/11/17/change-in-xs-terms-indicate-eu-researchers-will-get-api-access/?utm_source=substack&utm_medium=email

2. Quels autres leviers ?

Cette partie sera très largement à enrichir au fil des auditions, lectures et échanges collectifs.

Pour pallier aux carence précédemment détaillées, **plusieurs propositions ont été mises sur la table** aux travers de multiples publications et travaux de recherche :

- Dans la continuité des travaux de Célia Zolynski, le Conseil national du numérique "*défend l'idée de **consacrer un droit au paramétrage** qui permettrait aux utilisateurs de délimiter clairement leurs choix sans ingérence de la part de l'opérateur de plateforme. Ce droit pourrait concerner la recommandation et la modération de contenus ou encore la configuration de l'interface. Au-delà de l'intervention sur l'interface de la plateforme, il s'agirait de garantir l'accès à des API assurant l'existence d'applications tierces configurables, dont l'existence est aujourd'hui régulièrement menacée*"[38]. Cette proposition est également soutenue par la Commission nationale consultative des droits de l'homme (CNCDH)[39] Certaines plateformes camorcent d'ores et déjà un changement dans cette direction. Par exemple, la plateforme BlueSky propose aux utilisateurs de construire leurs propres algorithmes de recommandation et de les partager aux autres utilisateurs. Chacun est ainsi libre de paramétrer les contenus qu'il souhaite voir sur ce réseau social.

- Dans son bilan 2021, **l'Arcom formule une série de préconisation relatives à la transparence des algorithmes**, dont certaines seront de facto obligatoires suite à l'entrée en vigueur du RSN :
    - Préconisation n° 8 : **améliorer la transparence des processus et moyens dédiés à l'évaluation des outils automatiques** en publiant régulièrement, dans un espace dédié, des éléments sur les modifications qui ont été réalisées à l'aide d'analyses internes ou externes.
    - Préconisation n° 9 : **communiquer aux utilisateurs les critères utilisés pour la recommandation de contenus**, et ce, de manière personnalisée et contextuelle à l'aide d'outils et fonctionnalités accessibles directement sur le service.
    - Préconisation n° 10 : **instaurer une transparence accrue des politiques en matière de réduction de la visibilité des contenus**.
    - Préconisation n° 11 : **faire preuve d'une transparence plus importante sur le fonctionnement des outils de détection de contenus**, comptes et comportements utilisés à des fins de modération vis-à-vis du public et communiquer au régulateur davantage d'éléments permettant d'en évaluer la pertinence.

    - En matière d**'audit des algorithmes**, lors de la premier journée d'études de l'Arcom en novembre 2022, Marc Faddoul soutenait la nécessité de garantir la possibilité de **mener des audits adversariels** pour contrôler les effets des politiques des plateformes, au-delà des données auxquelles elles donnent accès (et dont elles ont donc le contrôle) et des audits indépendants obligatoires (qui ne seront pas publics)
    - Par ailleurs, pour améliorer la transparence et l'efficacité de l'audit algorithmique prévu à l'article 37 du DSA, lors de la journée d'études de Viginum en juin 2023, le collectif CheckFirst a indiqué souhaiter devenir auditeurs des plateformes et **proposer un outil de mesure public pour évaluer leur respect des règles édictées**.

Autres leviers potentiels :

---

[38] https://cnnumerique.fr/concevoir-sans-dark-patterns-webinaire-des-designers-ethiques. Voir également https://cnnumerique.fr/files/uploads/2022/Livres/CNNum-Votre-attention-sil-vous-plait.pdf

[39] CNCDH (2021). Avis sur la lutte contre la haine en ligne.

- Examiner en détail les possibilités de “désapprentissage” algorithmique pour sortir des bulles ;
- Faciliter signalement de contenus erronés, manipulateurs ou trompeurs et accroître la transparence du traitement de ces signalement ;
- Faciliter l’accès aux options permettant d’indiquer “ca ne m’intéresse pas” ou “je souhaite voire moins de ce type de contenus” ;
- Généraliser l’option permettant de voir rapidement pourquoi ce contenu m’est proposé.
- Lê Nguyên Hoang propose de valoriser les consensus : cette proposition s’inspire notamment d’une mesure prise à Taïwan avec la mise en place de la plateforme Pol.is mettant en avant les contenus consensuels - à l’inverse des plateformes mettant en avant les contenus clivants. Ils ont observé que cela a complètement changé la façon de produire l'information pour les créateurs de contenus et permet d’aller vers des conversations plus apaisées, nuancées et constructives.
-