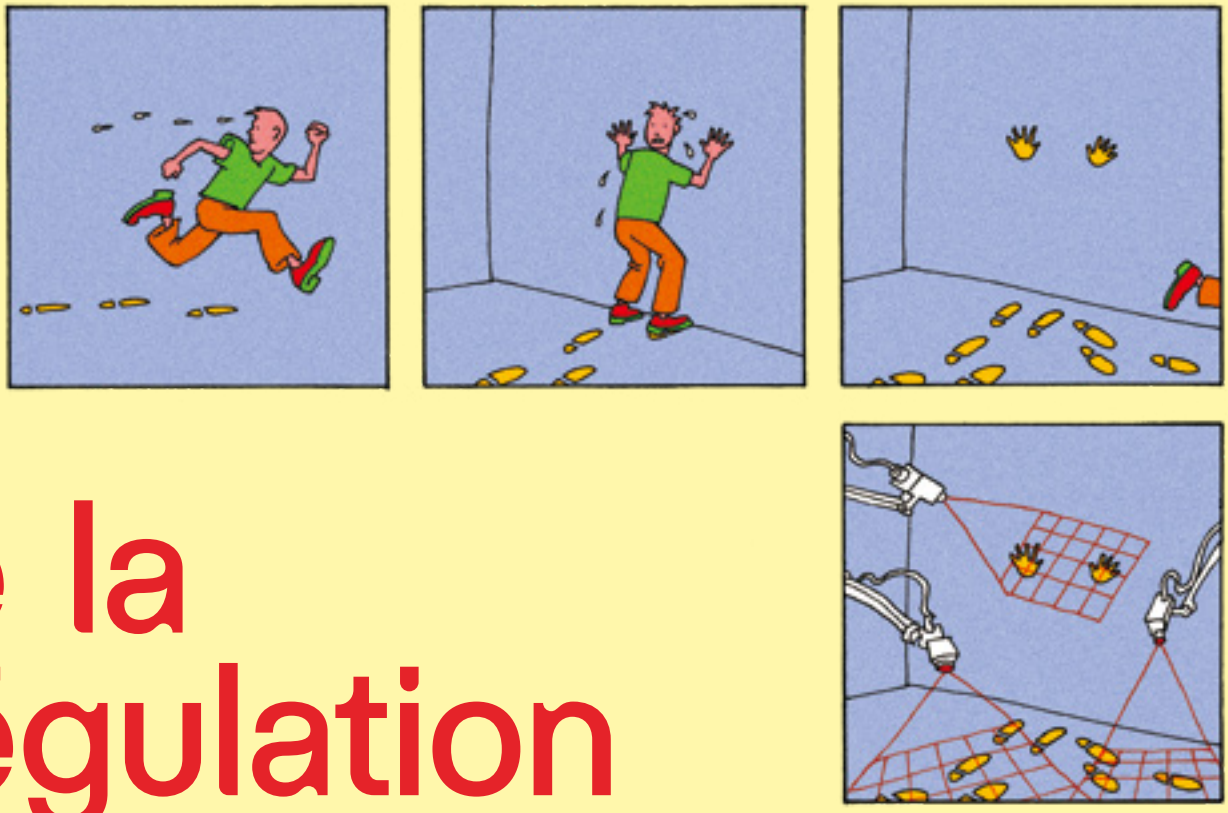
Nos traces numériques au service

Pedro Ramaciotti, chercheur CNRS et au médialab, dirige l'Observatoire européen de la polarisation (EPO). Ses domaines de spécialité sont les sciences sociales computationnelles, les systèmes complexes, les systèmes de recommandation, le *machine learning* et l'intelligence artificielle, l'analyse des réseaux sociaux en lien avec la politique.

Jean-Philippe Cointet est Associate Professor au médialab et dirige l'Institut libre des transformations numériques. Il est affilié au centre de recherche INCITE de l'Université de Columbia à New York. Ses rechervches se centrent sur la sociologie computationnelle et l'analyse

Tim Faverjon est doctorant au médialab, où il analyse les mécanismes d'apprentissage des algorithmes de recommandation.



de la régulation des plateformes

Par Pedro Ramaciotti,
Jean-Philippe Cointet et Tim Faverjon

La représentation géométrique du positionnement des partis politiques et des individus selon différentes dimensions, classique dans les études de politique comparée, n'a vu le jour que récemment dans l'analyse des données numériques. Les visualisations que proposent ici Pedro Ramaciotti, Jean-Philippe Cointet et Tim Faverjon découlent d'analyses réalisées à partir des traces numériques de comptes X/Twitter. Ces travaux ouvrent des pistes au régulateur pour prévenir le risque de profilage politique des utilisateurs des plateformes à leur insu.

La massification des échanges via les réseaux sociaux et la démocratisation des algorithmes d'apprentissage automatique, qui « calculent » les individus à partir de leurs traces comportementales, suscitent une défiance grandissante. Ces technologies, qui définissent la forme et les règles d'interaction au sein de l'espace public numérique, sont accusées d'accroître la polarisation des débats, d'encourager la prolifération de discours haineux, de propager de la désinformation (*fake news*), etc. De telles craintes renforcent l'attention portée aux moyens de régulation à disposition pour garantir les principes démocratiques.

Depuis le milieu des années 2010, l'Union européenne (UE) a élaboré un cadre régulatoire précurseur à travers une série de dispositifs légaux tels que le Règlement général sur la protection des données (RGPD), le *Digital Markets Act* (DMA), le *Digital Services Act* (DSA) et l'*Artificial Intelligence Act*. Deux d'entre eux, le RGPD et le DSA, sont censés enfin protéger les citoyens de l'Union européenne contre la collecte de données intrusives et les publicités qui utilisent des informations personnelles telles que les origines ethniques, les préférences sexuelles, la religion et les opinions politiques (article 26.3 du DSA, qui renvoie à la liste des catégories sensibles établies dans l'article 9.1 du RGPD). Linkedln a ainsi été épinglé, le 14 mars 2024, soit moins d'un mois après la mise en application du DSA, par la Commission européenne, qui soupçonne la plateforme d'utiliser des données sensibles (dont les préférences politiques) des utilisateurs pour les exposer à des publicités ciblées. Le DSA impose également aux opérateurs de plateforme,

On ne peut que se réjouir de voir l'Europe prendre un rôle de leader dans la protection des principes démocratiques en ligne.

en vertu de son article 34, de mesurer le risque que leurs services, dont les systèmes de recommandation et de modération, font peser sur la « liberté d'expression et d'information, y compris la liberté et le pluralisme des médias ».

On ne peut que se réjouir de voir l'Europe adopter un rôle de leader dans la protection des principes démocratiques en ligne. Il n'en est pas moins légitime de s'interroger sur l'efficience de ces outils juridiques. Le DSA interdit aux plateformes de faire du profilage politique à des fins publicitaires, mais de quels outils le régulateur dispose-t-il pour détecter ce type de profilage ? De même, une véritable responsabilité est donnée aux réseaux sociaux pour garantir la variété des opinions visibles en ligne. Or, les systèmes d'amplification, qui rendent les algorithmes si addictifs, sont également susceptibles de produire une vision tronquée ou biaisée des opinions. Passé ce constat, le problème qui se présente à la recherche et au régulateur est double : comment mettre en évidence et quantifier une telle déviation par

Collecte de données au Centre mondial de lutte contre l'idéologie extrémiste lors d'une visite officielle du président étatsunien Donald Trump à Riyad, Arabie saoudite, mai 2017.

rapport à un idéal pluraliste ? Comment mesurer la diversité des opinions exprimées sur un sujet donné ? Il faut aussi pouvoir délimiter l'espace dans lequel le respect de la diversité politique est souhaitable et définir le référentiel permettant de la mesurer : doit-on imaginer un indicateur idéologique sur le spectre droite-gauche ou envisager de la jauger dans d'autres dimensions attitudinales liées à des problématiques, parfois émergentes, telles que l'immigration, la mondialisation, les enjeux culturels ou environnementaux ?

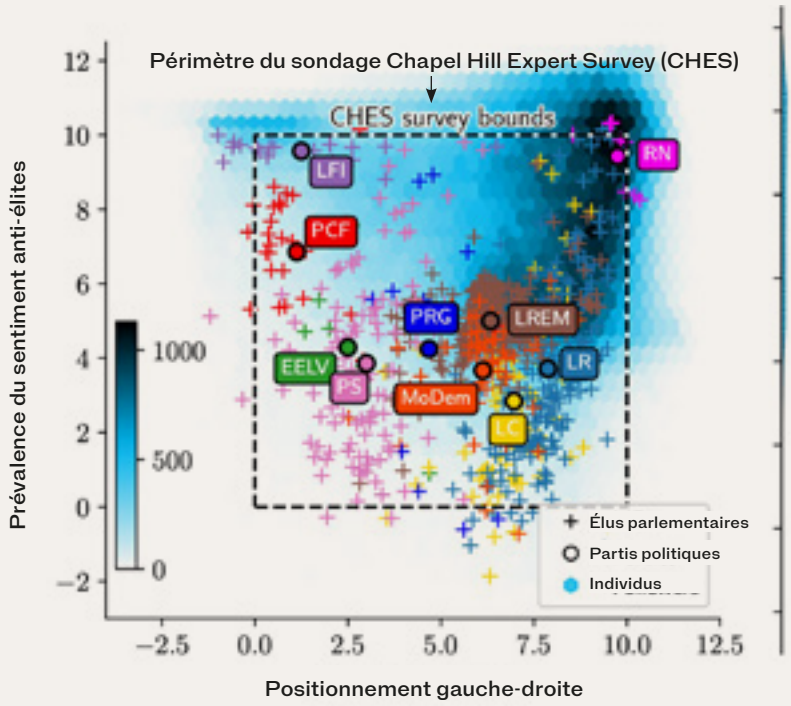
Mesurer l'opinion de larges populations à partir de leurs traces numériques

S'il est courant, dans les études de politique comparée, de recourir à une représentation géométrique pour positionner des partis ou des politiciens selon des axes prédéfinis, ce type de pratique n'a vu le jour que récemment dans l'analyse de données numériques. La nature de ces données, résultant généralement des traces comportementales laissées par les individus, dépend de chaque plateforme ; elles incluent typiquement des informations sur ce que les utilisateurs partagent, écrivent ou « aiment » (*like*). Elles sont particulièrement intéressantes lorsqu'elles sont produites par de grandes populations d'utilisateurs, car elles permettent d'extraire des conclusions sur les systèmes politiques nationaux à grande échelle avec une robustesse accrue.

L'estimation, sur la base de traces comportementales, des positions d'individus selon des dimensions ou échelles idéologiques[1] (opposant par exemple la droite à la gauche) ou des positions (pour ou contre) sur les différentes politiques publiques est une pratique relativement ancienne. Durant les années 1980, des travaux pionniers utilisaient des données de vote au Parlement pour positionner les législateurs sur des échelles idéologiques, l'intuition étant que les législateurs votant pour les mêmes lois étaient probablement très proches idéologiquement. A contrario, si leurs votes étaient rarement en accord, alors ils étaient à grande distance les uns des autres. Progressivement, l'ensemble des

[1] Une idéologie peut se décrire comme une structure d'attitudes adoptées face à un certain nombre d'enjeux du débat politique.

INFÉRENCE DU POSITIONNEMENT POLITIQUE DES INDIVIDUS SELON DEUX DIMENSIONS POLITIQUES, À PARTIR DE LEURS TRACES NUMÉRIQUES



Réalisé à partir des traces numériques de 400 000 utilisateurs de X/Twitter proches du débat politique et positionnés selon deux dimensions structurantes de la sphère politique française (opposition droite-gauche et sentiments anti-élites), calibrées avec les données du sondage CHES, ce graphique montre la forte prévalence du sentiment anti-élites chez les parlementaires, partis politiques et individus orientés à l'extrême droite et à l'extrême gauche. Les valeurs présentées en abscisses et en ordonnées (échelle de Likert) vont respectivement de 0 (extrême gauche) à 10 (extrême droite) et de 0 (pas de sentiment anti-elites) à 10 (très fort sentiment anti-élites).

Source : P. Ramaciotti *et al.*, « Inferring Attitudinal Spaces in Social Networks », *Social Network Analysis and Mining*, 13 (1), 2022, p. 14.

comportements dessinait un espace politique qui permettait de positionner finement chaque acteur dans un espace à une, deux, voire plusieurs, dimensions. Il en va de même aujourd'hui avec les traces numériques, qui peuvent trahir les préférences politiques des utilisateurs dès lors que l'on collecte les médias qu'ils retweetent ou les comptes de politiciens qu'ils suivent (pour ne mentionner que le cas de X/Twitter).

L'Observatoire européen de la polarisation[2] (EPO), coordonné par Sciences Po, s'attelle notamment à la tâche de mesurer l'opinion publique de larges populations (des centaines de milliers à plusieurs millions d'utilisateurs par pays) à partir de leurs traces numériques. Alors que les premiers travaux, fondés sur les traces des réseaux sociaux, visaient principalement à positionner des individus et des contenus sur des échelles opposant libéraux et conservateurs (en particulier pour l'analyse politique aux États-Unis), les recherches conduites au sein de l'EPO s'efforcent d'extrapoler ces travaux aux différents contextes nationaux présents dans l'UE. Des méthodes d'inférence statistique[3] sont élaborées à l'aide de diverses bases de données ayant servi à caractériser l'espace politique dessiné par les partis de chaque pays. Par exemple, les données de la Chapel Hill Expert Survey (CHES) sont utilisées pour positionner les partis politiques selon des dizaines de dimensions idéologiques ou sur des enjeux de politique publique qui structurent chaque contexte national : droite-gauche, UE, immigration, confiance dans les institutions et les élites, etc. Grâce à ces données d'expert, il est possible de valider et de calibrer les résultats obtenus par l'analyse des traces numériques et, surtout, d'étendre cette classification à l'échelle des partis politiques sur des populations très importantes.

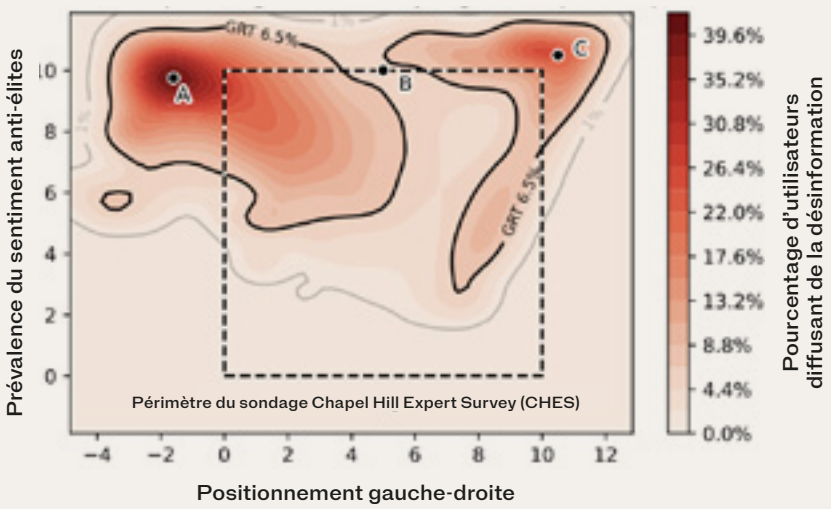
Mesurer les comportements et l'exposition en ligne en fonction des préférences politiques

Ces populations, parce que leur positionnement politique a été estimé selon les dimensions propres à leurs contextes nationaux et que ces estimations sont liées à leurs traces numériques (contrairement aux données d'enquête traditionnelle), peuvent constituer pour le régulateur une source privilégiée de métriques sur des questions ayant trait au risque de profilage politique des individus par les plateformes. Deux études, publiées respectivement en 2023 et 2024, l'illustrent : l'une porte sur la relation entre polarisation et désinformation en ligne, l'autre sur la recommandation algorithmique des contenus sur les plateformes sociales.

[2] L'Observatoire européen de la polarisation (European Polarisation Observatory, EOP) a été fondé par Sciences Po en 2021, en partenariat avec la London School of Economics, Bocconi University, Central European University, Hertie School et la Romanian National School of Political Sciences and Public Administration (réseau CIVICA) et en collaboration avec le consortium ExcellencES de France 2030 (CNRS, INRIA et Sorbonne Université). Il a pour mission d'étudier la polarisation dans des systèmes sociopolitiques européens multipolaires et multithématiques.

[3] Méthodes permettant d'inférer d'un groupe particulier les caractéristiques d'un groupe général, avec une probabilité d'erreur.

PRÉVALENCE DE LA DIFFUSION DE DÉSINFORMATION SUR X/TWITTER SELON LE POSITIONNEMENT IDÉOLOGIQUE DES INDIVIDUS



Réalisé à partir des traces numériques de 400 000 utilisateurs de X/Twitter proches du débat politique et positionnés selon deux dimensions structurantes de la sphère politique française (opposition droite-gauche et sentiments anti-élites), calibrées avec les données du sondage CHES, ce graphique montre que plus le sentiment anti-élites est élevé, plus la tendance à partager des *fake news* augmente. Les points A, B et C désignent respectivement les positions politiques de gauche, de centre et de droite parmi les utilisateurs ayant un sentiment anti-élites élevé. Le GRT (*global ratio threshold*) est le périmètre à l'intérieur duquel les utilisateurs partagent plus de *fake news* que la moyenne de la population.

Source : P. Ramaciotti *et al.*, « The Geometry of Misinformation. Embedding Twitter Networks of Users who Spread Fake News in Geometrical Opinion Espaces », *Proceedings of the International AAAI Conference on Web and Social Media*, 17, 2023, p. 730-741.

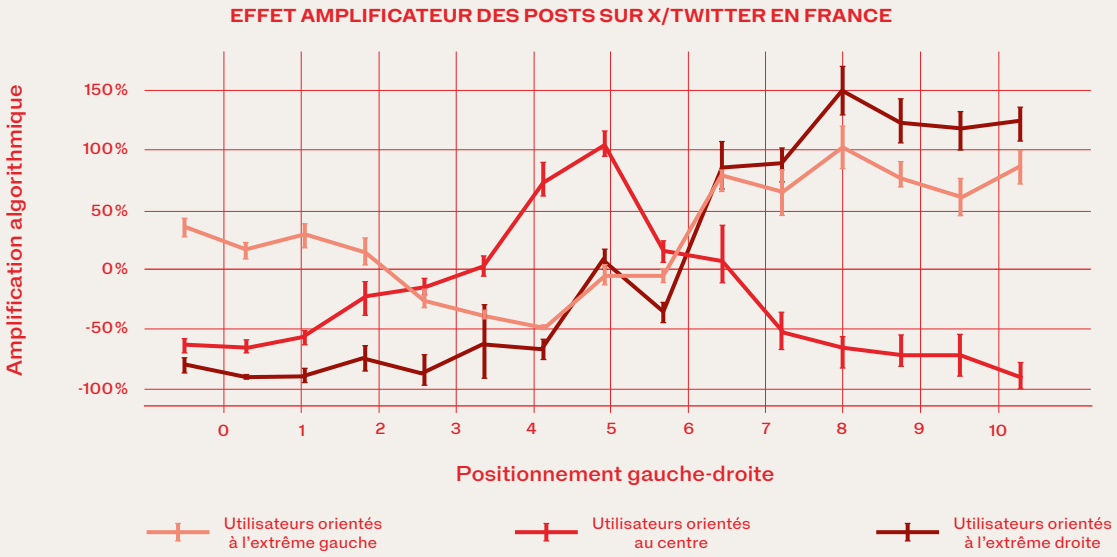
La lutte contre la désinformation en ligne est un des enjeux centraux de la modération et de la régulation des plateformes. Si l'on veut mieux la calibrer, il est indispensable de comprendre les déterminants du partage de *fake news*. Les recherches menées aux États-Unis ont montré que la désinformation était principalement diffusée par une faible portion de la population, située aux marges du spectre politique et, particulièrement, à l'extrême droite. Les populations produites dans le cadre d'EPO à partir de leurs traces numériques à l'échelle européenne permettent d'étendre aux différents pays des résultats déjà obtenus aux États-Unis, en prenant en compte les dimensions politiques propres qui structurent leur espace numérique. La meilleure illustration de ces résultats est l'étude de 2023 évoquée ci-dessus, qui analyse la désinformation circulant sur X/Twitter. Elle montre qu'en France le comportement de partage de *fake news* est largement déterminé par la position des comptes selon deux dimensions indépendantes : d'une part, l'axe droite-gauche, d'autre part (et peut-être avant tout), le sentiment anti-élites et la méfiance à l'égard des institutions qu'entretiennent certains comptes.

L'analyse de la recommandation algorithmique de contenus illustre , elle aussi, le défi auquel est exposé le régulateur. Pour se conformer à l'article 34 du DSA, les plateformes doivent évaluer l'impact de la recommandation algorithmique sur la pluralité et la liberté d'accès à l'information. Dans les pays où X/Twitter est la plateforme privilégiée des journalistes et des personnalités politiques, ce qui est le cas de presque tous ceux

La désinformation en ligne est un des enjeux centraux de la modération et de la régulation des plateformes.

d'Europe occidentale et d'outre-Atlantique, on peut imaginer les conséquences d'une amplification algorithmique ciblée qui privilégierait ou pénaliserait les messages et les contenus émanant d'un seul parti ou reflétant la perspective d'un seul camp politique.

Pour analyser ces questions, les chercheurs – auxquels l'article 40 du DSA attribue explicitement ce rôle – doivent disposer à la fois des données des recommandations des plateformes et d'une caractérisation politique du contenu recommandé et et de ses destinataires. Tel est l'objet de l'étude de 2024 évoquée ci-dessus sur la recommandation algorithmique, realisée en collaboration avec le CNRS (projet Horus) et menée à partir de populations numériques produites dans le cadre d'EPO. En mesurant conjointement les positions politiques des auteurs et des destinataires de messages recommandés, cette étude offre la première évaluation quantitative de la diversité politique des recommandations à laquelle sont exposés les acteurs de la twittosphère française.



Source : P. Bouchaud et P. Ramaciotti, « Auditing the Audits : Evaluating Methodologies for Social Media Recommender System Audits », *Applied Network Science Journal*, 9, 2024.

Ce graphique montre l'effet amplificateur des tweets sur X/Twitter résultant de l'application de la recommandation algorithmique sur la plateforme selon la position politique des auteurs et des lecteurs en France.

Elle montre clairement (voir figure ci-dessus) que les recommandations obéissent à une logique de ségrégation idéologique : les utilisateurs de gauche, du centre et de droite sont surexposés à des messages provenant de leur camp respectif. Autrement dit, les messages publiés par des amis partageant les mêmes opinions sont systématiquement amplifiés par l'algorithme. Seule exception, mais notable, l'algorithme amplifie également les messages d'utilisateurs d'extrême gauche chez les utilisateurs de droite, au détriment donc des contenus publiés par des centristes. La réciproque n'est pas vraie puisque l'on constate que les utilisateurs de gauche sont sous-exposés aux contenus émanant de la droite et du centre, bien qu'en moindre mesure pour ces derniers.

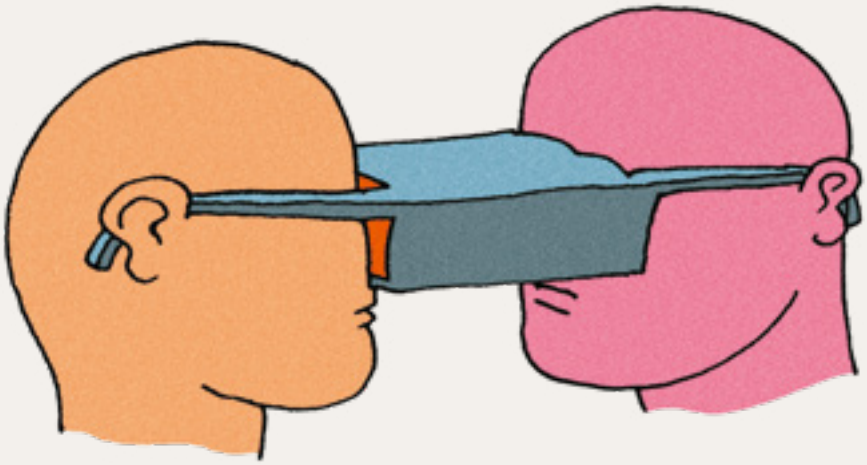
L'IA peut-elle générer des profils politiques par inadvertance ?

Les traces numériques des plateformes permettent ainsi de construire des ponts inédits entre informatique et études de politique comparée[4]. La question se pose de savoir si les algorithmes d'intelligence artificielle (IA) utilisés pour recommander des contenus sur les plateformes peuvent construire par inadvertance, dans leurs couches profondes, des profils politiques des utilisateurs.

Les technologies d'IA exploitent de grandes masses de données et produisent des modèles statistiques complexes pour calculer, par exemple, des prédictions ou des classements d'informations (qui alimentent les recommandations algorithmiques). Ces modèles ne sont cependant pas toujours compréhensibles ni explicables, d'où leur qualification fréquente de boîtes noires. Dès lors, on est en droit de pointer le risque que les algorithmes de recommandation internalisent dans leurs calculs, « à l'insu de leur plein gré », des profils politiques d'utilisateurs. Si oui, comment détecter ce phénomène, comment le mesurer et, éventuellement, comment s'en prémunir ?

Ces questions se justifient pour deux raisons. Premièrement, la création de profils au sein des modèles de l'IA constituerait une violation de l'article 26 du DSA et signifierait, en pratique, que les plateformes se dégageraient de leurs responsabilités en s'abritant derrière l'opacité des modèles. La détection de ces profils dans les modèles d'IA pourrait également permettre d'empêcher des violations intentionnelles, mais furtives, de l'article 26 par les plateformes. Par exemple, si l'opérateur d'une plateforme est convaincu que son modèle d'IA fournira des publicités politiques pertinentes à ses utilisateurs (en anticipant quel contenu sera montré aux utilisateurs d'un bord politique), sans avoir à l'inscrire explicitement dans la conception de son modèle d'IA, il pourra proposer la publicité politique ciblée comme un

[4] Tel est l'objectif du projet « AI-Political Machines » (AIPM), financé par le Project Liberty Institute à Sciences Po, qui utilise les populations dotées de positions politiques produites dans le cadre d'EPO.



service tout en prétendant que le profil politique des utilisateurs reste inconnu de la machine. Deuxièmement, les efforts destinés à modérer les phénomènes négatifs dus à la diversité politique des contenus consommés (par exemple, la polarisation exacerbée) soulèvent généralement des problèmes complexes de normativité : quel degré de diversité de contenus faut-il imposer aux utilisateurs ? Qui doit le mesurer et qui doit l'imposer ?

Il est par ailleurs permis de penser que les modèles d'IA soient capables d'effacer sélectivement les informations susceptibles de trahir les préférences politiques d'un individu. Dès lors, ne pourrait-on concevoir des systèmes de recommandation aveugles à la politique, conformes à la législation, mais restant pertinents pour l'utilisateur ? Se doter de la capacité de cartographier l'espace politique que les traces numériques laissent deviner est un enjeu clé pour répondre à cette question. Et il est crucial à cet égard que les données propres aux plateformes numériques soient largement interrogeables par la recherche.

Les recommandations algorithmiques obéissent à une logique de ségrégation idéologique : les utilisateurs sont surexposés à des messages provenant de leur camp politique respectif.

RÉFÉRENCES

→ Bouchaud P. et Ramaciotti P, «Auditing the Audits. Evaluating Methodologies for Social Media Recommender System Audits», *Applied Network Science Journal*, 9, 2024.

→ Ramaciotti P., Cointet J. P., Munoz Zolotoochin G., Fernandez Peralta A., Iniguez, G. et Pournaki A., « Inferring Attitudinal Spaces in Social Networks », *Social Network Analysis and Mining*, 13 (1), 2022, p. 14.

→ Ramaciotti P., Berriche M., et Cointet J. P., « The Geometry of Misinformation. Embedding Twitter Networks of Users who Spread Fake News in Geometrical Opinion Espaces », *Proceedings of the International AAAI Conference on Web and Social Media*, 17, 2023, p. 730-741.